# Prediction of Land Cover in Continental United States Using Machine Learning Techniques

A Thesis
Presented to
The Academic Faculty

by

**Yashika Agarwalla**

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

School of Civil & Environmental Engineering
Georgia Institute of Technology
May 2015

# Prediction of Land Cover in Continental United States Using Machine Learning Techniques

Approved by:

Dr. Marc Stieglitz, Advisor
School of Civil & Environmental Engineering
*Georgia Institute of Technology*

Dr. Greg Turk
College of Computing
*Georgia Institute of Technology*

Dr. Mustafa Aral
School of Civil & Environmental Engineering
*Georgia Institute of Technology*

Date Approved : April 19, 2015

*Mom & Dad,*

*thanks for holding me through this.*

# ACKNOWLEDGEMENTS

First and foremost, I would like to extend my heartfelt gratitude to my advisor, Dr. Marc Stieglitz for all his unconditional guidance without which I wouldn't have been anywhere close to completing this. I am very grateful to my lab mate, Felipe Dias for being ever ready to help me through this and always answering my queries, how so ever silly and redundant they were. A big shout goes out to my parents for sending me miles away to attend a school as magnanimous as Georgia Tech and to all my friends who had my back through all my tough times. On a concluding note, I would like to acknowledge my jury for taking the time of their busy schedules for reading through my thesis and for their indispensable and wise remarks. I'm highly obliged to have received all your help and no amount of acknowledgment would parallel all your help and support.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# SUMMARY

Land cover is a reliable source for studying changes in the land use patterns at a large scale. With advent of satellite images and remote sensing technologies, land cover classification has become easier and more reliable. In contrast to the conventional land cover classification methods that make use of land and aerial photography, this research uses small scale Digital Elevation Maps and it's corresponding land cover image obtained from Google Earth Engine. Two machine learning techniques, Boosted Regression Trees and Image Analogy, have been used for classification of land cover regions in continental United States. The topographical features selected for this study include slope, aspect, elevation and topographical index (TI). We assess the efficiency of machine learning techniques in land cover classification using satellite data to establish the topographic-land cover relation. The thesis establishes the topographic-land cover relation, which is crucial for conservation planning, and habitat or species management. The main contribution of the research is its demonstration of the dominance of various topographical attributes and the ability of the techniques used to predict land cover over large regions and to reproduce land cover maps in high resolution. In comparison to traditional remote sensing methods such as, aerial photography, to develop land cover maps, both the methods presented are inexpensive, faster. The need for this research is in synergy with past studies, which show that large-scale data, processing, along with integration and interpretation make automated and accurate methods of change in land cover mapping highly desirable.

# CHAPTER I

# INTRODUCTION

Land cover is the term used to describe man-made and natural features present on Earth's surface [1]. It is a reliable source for studying the change and patterns in land use at a large scale [2, 1]. Understanding the spatial distribution of various vegetation types which is affected by climate and topography has been of ecological importance [3, 4, 5, 6]. In the face of today's changing environment and urbanization, changes in land use or land cover (LULC) can be an essential tool for planning utilization of natural resource management [7]. Up-to-date LULC information is of critical importance to planners, scientists, resource managers, and decision makers [8]. Baatuuwie and Leeuwene (2001) developed a suitable method of mapping the different forest stand types using GIS and remote sensing in Ghana for monitoring and managing the forest resources in this area. Information on LULC spatial distribution can be useful in modeling earth's systems [9]. LULC changes lead to a various changes such as loss of biodiversity, desertification, and climate change, etc [2]. Land cover changes undoubtedly influences species distributions. Habitat loss alters the ecological processes which in turn can lead to species extinctions [10, 11]. The spatial distribution of trees and shrubs species shown through a land cover map is helpful in understanding forest status [12]. A recent study by Shirley et al.(2013) uses Landsat data to predict occurrence of bird species in forest landscapes of western Oregon, USA [11]. Forest distribution varies all around the globe and can range from tropical rain forests to cold and dry taiga with different structures, distribution and compositions [12]. LULC information can be obtained from various Remote Sensing techniques.

Remote Sensing is the technique used to acquire information for detection and classification of objects on Earth. Traditional methods of detecting land cover include using field

and aerial photography, both of which prove to be very tedious and expensive [2]. In recent decades, with availability of easily accessible satellite data on platforms such as Google Earth Engine, satellite data have now become one of the primary sources for obtaining information about the vegetation on the Earth's land surface and with it various unconventional methods are being researched for land-cover classification [13, 12, 14]. The UN Conference on Environment and Development at Rio de Janeiro and Kyoto also cited satellite imagery as a reliable and promising method for detailed mapping and monitoring of forests resources [15]. The Landsat 7 satellite captures six bands of the visible and infrared spectrums at high resolutions of 30 m every 16 days [16, 17, 11]. As the capabilities of remote sensing based mapping and monitoring programs improve, attempts have been made to characterize large spatial regions [18]. A master's thesis research work carried out at University of Lethbridge (2013) studies the effects of primary topographic variables (such as, slope, elevation, and aspect) and compounded topographic variables (such as, topographic wetness index and solar radiation) on land cover [19]. Over the past two decades, use of Machine Learning algorithms have commonly outperformed conventional remote sensing techniques [20].

Machine Learning (ML) is a field of computer science where data is transformed into meaningful action or information without having to program it explicitly. ML includes various algorithms with the ability to recognize recognize data patterns through repeated learning techniques [21]. It also closely related to the field of data mining where patterns are extracted from huge data sets. ML employs the use of techniques such as statistics and optimization to summarize the raw data in form of a model or a relation which can be equations, trees, logical if or else sequences and clusters of data [22, 23]. Decision making tools such as linear models, decision trees, clusters, etc. are commonly used in ML to predict the response or output value based on several input variables [23]. A variety of ML techniques have been used for land cover classification, such as unsupervised clustering algorithms, parametric supervised algorithms, decision trees, random forests, neural networks, etc [24, 25, 26, 27, 20, 13]. The literature reviews show that decision tree based algorithms

and their variants have been used due to their simple interpretation, high classification accuracy, and ability to characterize complex interactions among attributes [28, 29, 30]. Kalbi et al. (2014) discusses the efficiency of using Boosted Regression Trees, Random Forests (RF) and Classification and Regression Trees (CART) for forest type mapping using satellite data [12]. Their study demonstrates that Random Forests and Boosted Regression Trees have higher prediction accuracies in comparison to Classification and Regression Tree. This thesis makes use of the boosting algorithm for BRT which creates a highly accurate learner by combining several weak learners [31]. The various criterion for evaluating ML algorithms for land cover classification using satellite data was researched by DeFries and Chan(2000) [13]. In Japan, a research involved using four different classifiers namely; CART, RF, Decision Trees with boosting and Decision Trees with bagging which showed that RF and Decision Trees with boosting to be most reliable and efficient methods for land cover mapping [32].

The two machine learning techniques used in this study are Image Analogy and Boosted Regression Trees. Although, ML classifiers are an improvement to the conventional remote sensing techniques, research is still needed to assess their usefulness when compared with each other [13]. The objective of this thesis is to use ML techniques to create land cover images from satellite images and data and enable scaling up the land cover prediction to neighboring regions as well. Topography is an essential attribute that contains geological, geomorphological and climatic information of a region [19]. Its importance is based on the fact that variability in relief and topographic variables such as elevation, slope, and aspect influence different land cover types [33, 6, 3]. Land cover variation is according to topographical differences which leads to varying vegetation types; this favors a range of habitat conditions that fosters biodiversity. Thus, through this thesis, we establish the topographic-land cover relation which is very essential in conservation planning, and habitat or species management [34, 35, 36, 19]. The topographical features selected for this study include slope, aspect, elevation and topographical index (TI) which is also referred to as wetness index as they have been proven to be powerful factors determining land cover distribution [6, 33, 35, 19]. These topographic variables can be measured with a digital elevation model

(DEM) obtained from the Google Earth Engine, which is a grid of data with each cell containing an elevation, slope and aspect value [37, 19]. A greater diversity in land cover types is an indicator to a to an area with high biodiversity which is supported by a stable and productive ecosystem [6, 35, 38, 39, 40, 41, 19]. This study carried out for regions in continental United States uses small scale Digital Elevation Maps and it's corresponding land cover image obtained from Google Earth Engine and hence enable scaling up predictability up to almost five times. In this thesis, we assess the efficiency of ML techniques in land cover classification from satellite data and understand the relation between topography and land cover. This thesis explains the algorithms, analyzes the performance and discuss the technical issues of the two ML techniques being used here; Boosted Regression Tree and Image Analogy.

Image Analogy (IA) is a way of creating a new image from a target image by applying the same relation that was present in training pair of images [42]. IA is a ML technique using computer graphics and it has been used for various purposes such as traditional image-filters, super resolution, artistic filters and texture transfer [42]. Here, IA has a new found approach of recreating photorealistic images of regions in continental United States using small scale Digital Elevation Model and Landsat images as reference image and transferring the relation over a larger area. The most intriguing feature of IA is it creates photorealistic images but it lacks the ability to predict the most dominating features within a region due to which Boosted Regression Trees were brought into the picture.

Boosted Regression Trees (BRT) is a ML technique that improves the predictive performance of a weak single regression tree by combining several regression trees. BRTs can be stated as an ensemble of regression trees and boosting techniques used to predict the relative importance of various variables in a model. Before moving on with BRT it is important to understand regression trees which is essentially its building block. *Figure* 1.1 is a schematic representation of a regression tree which is a tree like structure comprising of branches and leaves. In the figure, each branching point($t_1$, $t_2$, $t_3$, .... $t_4$) represents a decision variable ($X_1$, $X_2$) or a test condition following which the input space is divided

into branches leading to the output which are the leaves ($R_1$, $R_2$, $R_3$, .... $R_5$). The output or response variable (forest or non-forest) is a function of input variables. However, due to its poor prediction ability, boosting techniques are used. Boosting is a technique of improving models predictive accuracy by reducing the loss function (i.e. the difference between the ground truth and the predicted value) and it can be of squared error or gradient loss function types. It was first started in 1996 at AT&T Labs by Freund and Schapire [43]. The technique combines several regression trees to reduce the residuals or loss function($L$) i.e. the difference between the original and the predicted value.

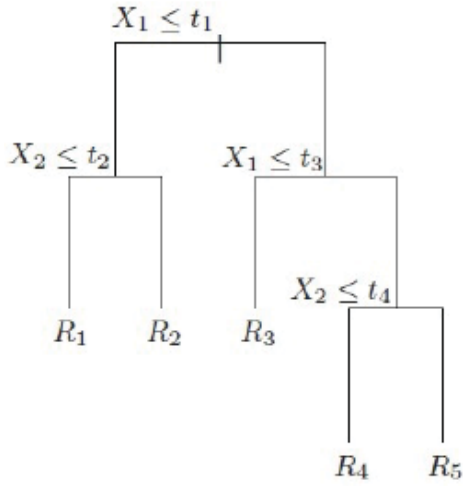



**Figure 1.1:** A schematic representation of a Regression Tree

BRTs have found to be of purpose in pattern prediction by ecologists to and also in several cases to find most influential variables. This section first starts with explaining the regression tree algorithm, then moves to techniques used in gradient tree boosting and finally explains the results. The popularity of BRT arises from their intuitive and easy to visualize capabilities [28].

The techniques have been explained in more details in Chapter II followed by Chapter III which talks about the various data sets that have been used. Chapter IV explicitly gives the detailed methodology followed. The results of IA and BRT have been individually presented in Chapter V and VI respectively. Finally, Chapter VII gives a comparison of the two ML techniques followed by discussions and conclusions.

# CHAPTER II

# TECHNIQUES USED

Two techniques used for prediction of vegetation cover viz. Image Analogy and Boosted Regression Trees, both of which have been explained in details in sections below.

## *2.1 Image Analogy*

Image Analogy(IA) is a method of synthesizing a new image (B') from a target image (B) using the analogy that is generated using source image A and A' as a training pair. It can be stated in a simpler way by saying that, IA creates an image B' which relates to B in the same way as the image A' relates to A [42] (*Equation* 2.1*a*). The image pair A and A' are registered images of same dimensions. [42].

$$A : A' :: B : B' \tag{2.1a}$$

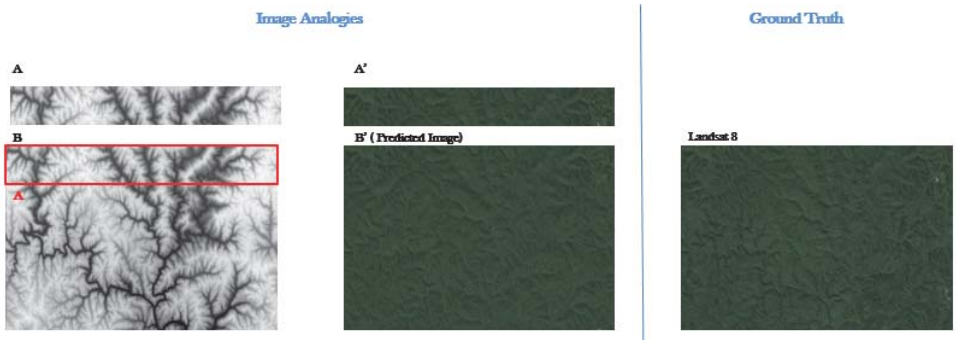



**Figure 2.1:** Image Analogy Representation

The IA algorithm uses the Approximate Nearest Neighbor (ANN) approach which has been depicted in *Figure* 2.2. The source pixel, $p$ in source pair of images A and A' is referred to as $A(p)$ and $A'(p)$ and $q$ is the target pixel in target pair of images $B$ and B'

referred to as $B(p)$ and $B'(p)$. The IA algorithm as schematically represented in *Figure* 2.2 proceeds by finding the best match of pixel $q$ in image $B$ from pixel $p$ in image $A$ and then a pixel $q$ is synthesized in image B'. The synthesis of B' takes place over a series of resolution levels from coarser to finer where in $l$ is the final resolution level and $l-1$ is a level lower resolution. The program usage has been explained in *Table* 2.1.

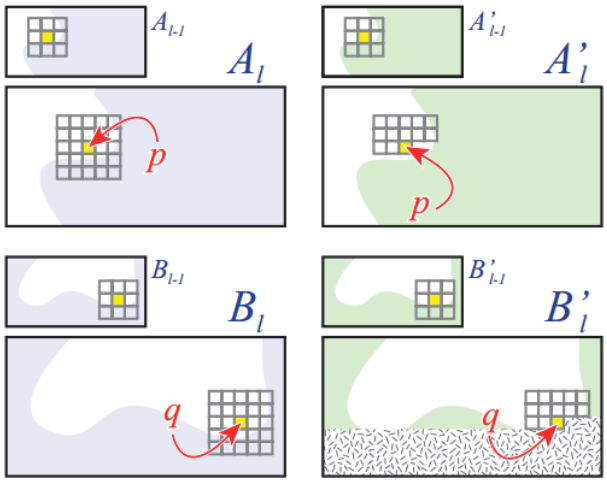



**Figure 2.2:** Image Analogy Algorithm Representation using ANN [1]

## 2.2 *Image Quality Assessment Metrics*

Image Quality Assessment methods are algorithms used to evaluate the image quality. The Image Quality Assessment methods used here to obtain the degradation of the predicted image created using IA from the original image. The metrics are explained as under and the results obtained will be discussed in the following section.

**Mean Squared Error (MSE)** : MSE is the mean squared difference between the original and the distorted image which is calculated by adding the squares of the differences in the pixels and dividing by the total number of pixels [45] (*Equation* 2.2).

[1]Source : A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* pp. 327–340 ACM 2001

[2]Source : "Chris Tralie : Image Analogies,"

**Table 2.1:** Parameter definitions in Image Analogy taken from Chris Tralie's Image Analogy [2]

| Parameter | Purpose |
|---|---|
| **A, A', B, B'** | A and A' are the input images of same size with some relation between each other. B is also an input image, from which B' is synthesized such that A $\rightarrow A' \sim B \rightarrow B'$. |
| **-mask_color r g b** | Used to specify the mask color (r, g, b) used for the hole filling/image inpainting feature. r, g, b are in the range [0, 1.0]. |
| **bruteForce** | Use an exhaustive search over all feature vectors instead of ANN. By default, brute force is not used. |
| **luminanceRemapping** | To perform luminance remapping or not. |
| **noGaussianLuminance** | Option to scale down neighborhoods around a pixel by a gaussian (so that neighboring pixels that are closer get emphasized more in the feature vector). False by default. |
| **steerableFilters** | Inclusion of steerable filters at the end of each feature vector in addition to luminance neighborhoods |
| **kappa** | Specifies the coherence parameter. This is set to zero by default. |
| **ANNEps** | The allowed error of the Approximate Nearest Neighbors library (default 1.0) |
| **levels** | Multiresolution synthesis levels. |
| **outputPyramid directory/previx** | If this parameter is specified, the entire gaussian pyramid of B' will be outputted to the directory specified. |
| **verbose** | Outputs each scanline as it is finished (useful to make sure program isn't hanging at some scanline). |

$$MSE(A,B) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{2.2}$$

Unlike MSE which does not take into consideration human vision, the next two methods are based on Human Vision System (HVS). HVS is a complex way of comparing images based of perception to the human eye taking attributes such as luminance, contrast, texture, brightness into consideration separately rather than taking it all as a whole [46]. MSE is a pixel to pixel method and does not take relation between pixels into account and is averaged as it is computed on a window of 8 x 8 pixels sliding over the whole image. Two of the HVS metrics used have been described below:

**Universal Image Quality Indes (UIQI)** : UIQI is different from MSE and PSNR as it takes into account human visual perception. UIQI is a product of luminance, contrast and structural comparisons where $\mu_a$, $\mu_b$ are the mean of original and predicted images, $\sigma_a$, $\sigma_b$ are the standard deviations of original and predicted images and $\sigma_{xy}$ is covariance of the two images [46, 47]($Equation$ 2.3, 2.4, 2.5).

$$l(a,b) = \frac{2\mu_a\mu_b}{\mu_a^2 + \mu_b^2} \tag{2.3}$$

Equation 2.3 measures how close the luminance between the two images, a and b and it ranges within [0,1].

$$c(a,b) = \frac{2\sigma_a\sigma_b}{\sigma_a^2 + \sigma_b^2} \tag{2.4}$$

Equation 2.4 is a measure of how similar the contrasts of the two images are.

$$s(a,b) = \frac{2\sigma_{xy}}{\sigma_a + \sigma_b} \tag{2.5}$$

Equation 2.5 computes the degree of correlation between the two images and falls within the range [-1,1].

UIQI is a product of $l(a,b),\ c(a,b),\ s(a,b)$.

**Structural Similarity Image Metric (SSIM)** : SSIM is an improvement over UIQI by Wang et al as constants, $C_1$, $C_2$, $C_3$ are introduced here to prevent the division of terms by zero and $\alpha$, $\beta$ and $\gamma$ are the weights of luminance, contrast and structure terms respectively [48](*Equation* 2.6, 2.7, 2.8, 2.9). SSIM value ranges from -1 to 1, where the value of 1 is attained for a perfect match.

$$l(a,b) = \frac{2\mu_a\mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \tag{2.6}$$

$$c(a,b) = \frac{2\sigma_a\sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \tag{2.7}$$

$$s(a,b) = \frac{2\sigma_{xy} + C_3}{\sigma_a + \sigma_b + C_3} \tag{2.8}$$

Where $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$, two constants to stabilize the division with weak denominator [49]

L is the dynamic range of the pixel-values ( $2^{\#bits\ per\ pixel} - 1$)

$k_1 = 0.01$ and $k_2 = 0.03$

$$SSIM(a,b) = l(a,b)^{\alpha}.c(a,b)^{\beta}.s(a,b)^{\gamma} \tag{2.9}$$

**Feature Similarity Image Metric (FSIM)** : FSIM is based on Phase Congruency(PC) and Gradient Mean(GM) comparison between the reference and the predicted image. These two are low-level features that complement each other.

The IQA techniques used here are all full reference methods as the original image from Google Earth Engine was available for comparison.

## *2.3 Boosted Regression Trees*

A regression tree predicts the value of a response by following a flow chart pattern where at each branching point a decision is made depending upon which it further branches [50]. As shown in *Figure* 1 a regression tree works by splitting the input variables (referred as attributes from here on) into various non-overlapping subspaces $R_1$, $R_2$,... $R_M$ such that union of these regions comprise the whole set of attributes. A tabular representation of the attributes and response variables has been shown in *Table* 2.2.

**Table 2.2:** Representation of input & output variables

| Attributes | | | | Response |
|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | ... | $x_{1m}$ | $y_1$ |
| $x_{21}$ | $x_{22}$ | ... | $x_{2m}$ | $y_2$ |
| ⋮ | ⋮ | | ⋮ | ⋮ |
| $x_{n1}$ | $x_{n2}$ | ... | $x_{nm}$ | $y_n$ |

Here the input($x$) consists of $m$ attributes each having $n$ instances and each attribute can be grouped into vectors $X_1$, $X_2$ ... $X_n$. $y_1$, $y_2$ ... $y_n$ is the output or response variable predicted by the regression tree.

The first step in building a regression tree is to choose an attribute, $X_1$ and within it a branching point, $\theta$ such that it divides the input space into two regions $X_1 < \theta$ and $X_1 > \theta$. Recursive branching is carried out until a stopping criterion is reached which could be a size limit, all branches on partitioning have the same class and no features are left to be distinguished [22]. The branching point is chosen to minimize the squared error loss between the sum of the squares of the differences of the observed value and mean of

responses created in each region from the division at branching point [50] (*Equation* 2.10). Since none of the regions are overlapping, each $x$ would exactly belong to only one region, viz. $R_1$, $R_2$ or $R_5$. The terminal point would give the predicted value of $y$ or the class.

$$\Sigma_{x_i \epsilon R_1} \ (y_i - \beta_1)^2 + \Sigma_{x_i \epsilon R_2} \ (y_i - \beta_2)^2 \leq \Sigma_{x_i \epsilon R_1'} \ (y_i - \beta_1')^2 + \Sigma_{x_i \epsilon R_2'} \ (y_i - \beta_2')^2 \qquad (2.10)$$

Where $\beta_1$, $\beta_2$, $\beta_1^{'}$ & $\beta_2^{'}$ is the mean of the responses within the regions $R_1$, $R_2$, $R_1^{'}$ & $R_2^{'}$ respectively.

Boosted Regression Tree (BRT) is combination of two steps namely; Forward Stage Additive Modeling (FSAM) and Gradient Descent Optimization [51] which will be been explained in detail below. New regression trees are developed from residuals and added to previous models to reduce the error [52]. BRT with starts with one regression tree, then proceeds with recursive addition of more regression trees to reduce the loss function. This technique was introduced in 1999 by Jerome H. Friedman [53]. The differences between BRTs and regression trees are BRT employs the technique of boosting by using several trees for prediction as against a regression tree which uses just one to find several important rules rather than just finding one rule. Unlike regression tree, BRTS are combinations of several regression trees with improved predictive performance, the ability to fit complex relations, and accommodate missing data.

The BRT can be expressed as a function, $(f(x))$ that maps the relation from $x \rightarrow y$ (*Equation* 2.11$a$) where in $g(x)$ is a regression tree with attributes($x$) and branching point($\theta$). A decision tree function can be represented as in *Equation* 2.11$b$ where $J$ is the total number of regions in the tree, $\gamma_j$ is the predicted value assigned to the region, $R_j$ and $I()$ takes the value of 1 if $x$ belongs to the region and 0 otherwise. At each iteration($i$), $f(x)$ is optimized using a minimum loss function to attain optimum values of $a_i$ and $\theta_i$ (*Equation* 2.11$c$). The loss function is defined as the square of the difference between the given predictor value and the value given by $f(x)$. In order to find the value of $\theta$ at which the loss function is minimized gradient descent optimization is used. The technique finds

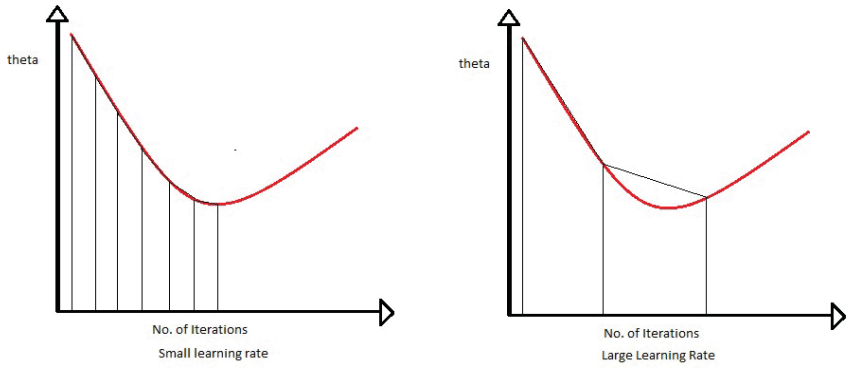



**Figure 2.3:** Optimization of $\theta$ depending on $\rho$

the value of parameter $\theta_i$ at each iteration such that the loss function attains its global minima by taking partial derivative of the loss function. The optimized value of parameter in previous iteration was $\theta_{i-1}$, then the updated value $\theta_i$ is obtained (*Equation* 2.11*d*) where $\rho$ is the learning rate and $\partial L/\partial\theta|_{\theta_{i-1}}$ is the partial derivative of loss function at $\theta_{_{i-1}}$. The learning rate, $\rho$ is the contribution of the variable to the function which ranges from 0 to 1. The lower the learning rate, the better is the optimization in attaining global minima. As shown in *Figure* 2, for lower learning rate the value is optimized in small steps to reach minimum value but for larger learning rate the change is so large that it might fail to reach a minimum value. At every iteration, the optimized function is added to the function from previous iteration. The final BRT shown in *Equation* 2.11*e* can be thought of as a linear combination of several trees where each term of the function represents one tree [54, 55, 28].

$$f(x) = \sum_{i=1}^{I} a_i g(x, \theta_i) \tag{2.11a}$$

$$g(x, \theta) = \sum_{j=1}^{J} \gamma_j \; I(x \epsilon R_j) \tag{2.11b}$$

$$(a_i, \theta_m) = arg \; min \; L \; (y, f_{i-1}(x) + a \; g(x, \theta)) \tag{2.11c}$$

$$\theta_i = \theta_{i-1} + \rho \partial L / \partial \theta|_{\theta_{i-1}} \tag{2.11d}$$

$$f_i(x) = f_{i-1}(x) + a_i g(x, \theta_i) \tag{2.11e}$$

# CHAPTER III

# DATA & SOFTWARE USED

## *3.1 Google Earth Engine*

Google Earth Engine is a planetary-scale platform by Google which brings together satellite imagery from all around the globe and makes it available online to a wide gamut of people enabling them to make use of this vast collection of datasets and carry out interesting research and studies [56]. It contains various datasets and tools such as Landsat Imagery, NDVI, surface reflectance which dates back to almost 40 years. The following datasets have been used from Google Earth Engine:

**Elevation**: USGS National Elevation Dataset 1/3 arc-second

**Landsat** : Landsat 8/32 TOA Reflectance Composite

**Classified Image** : Using add points and Train a Classifier Tool

**Slope & Aspect** : Using Slope and Aspect under Add Computation Tool

## *3.2 Grass GIS*

GRASS GIS (Geographic Resources Analysis Support System) is an open source, freely available geographical information system (GIS) used to handle raster, vector data and it also finds use in image processing applications [57]. Grass GIS was used to import all the datasets downloaded from Google Earth Engine. Once the importing was completed two separate datasets for training and testing were created. A training data set was almost 1/5th of the entire region and this region was decided in a manner such that most of the elevation gradient was captured in the training dataset. Once these regions were defined then the training and test data sets were exported as .csv files. For usage in IA, the regions of images were saved as .png files to be used as training and test pairs. Grass GIS was also used to import the results of predicted data from IA and BRT and compare them to the originally classified image obtained from Google Earth Engine.

## *3.3 R*

R is a free, open sources programming language, developed by Ross Ihaka and Robert Gentleman at University of Auckland. It is used for statistical computing and graphics [58]. The BRTs were modeled in R version 3.1.0 using the gbm package version 2.1. The training and test files were read in as .csv files and the output binary classification i.e., forest or non-forest (0 *or* 1).

# CHAPTER IV

# METHODOLOGY

## *4.1 Image Analogy*

The prediction of vegetation cover using IA is carried out in two phases, first being the training where the DEM image of a smaller region(A) and its corresponding visual image(A') obtained from Google Earth Engine is used as training pair and second being the prediction where the analogy obtained between the two is applied to a DEM image of a larger region(B), almost 5 times the size of $A$ and containing the smaller region $A$ to obtain a predicted image **B'** (*Figure* 2.1). The images A and B are obtained from the USGS National Elevation Dataset 1/3 arc-second dataset available on Google Earth Engine. The image A' and the original image against which B' is compared is obtained from the Landsat TOA Percentile Composite [56]. After having chosen the desired region, these layers were downloaded at a preferred resolution and using UTM coordinates. The UTM zone map was used to decide the zone in which the region lies [59]. The selection of the region to be used in training pair is done in a manner such that the maximum elevation gradient is captured in the image A. The parameter values used are as follows; kappa : 30 and level : 4. Once, IA were carried out, the five image quality Metrics discussed Chapter II were used to evaluate the image quality. The methodology used for IQA has been explained in following paragraph.

The original images used as reference images were extracted from Google Earth Engine which were compared with the predicted images generated using IA. All these metrics require for the image to be in grayscale to be compared, thus the images were imported into Matlab and converted into grayscale scale RGB2gray and then the comparisons were made using the five metrics. *Appendix* has the codes along with the sources for the IQA metrics used.

## *4.2 Boosted Regression Trees*

The BRTs used here has four attributes namely; elevation, slope, aspect, topographical index (TI) and a class Forest(1) or Non-Forest(0). Each pixel in an image of a selection region has the aforementioned four attributes and a class label against it. Elevation is obtained from the USGS National Elevation Dataset 1/3 arc-second dataset available on Google Earth Engine and then the layer was downloaded as a GeoTiff file with UTM coordinates. The slope and aspect for a region is calculated using built-in functions in Google Earth Engine. Once the slope and aspect were calculated, the layers were downloaded as explained above for elevation. In order to compute the Class of every point, sample regions were selected to demarcate the two classes were created viz.; forest and non-forest and color coded green and red respectively. Then using the point drawing option a few points were created for each class using the underlying satellite image as reference. Then the Fast Naive Bayes classifier was applied to create a binary classified image which was downloaded as a GeoTiff file. Then all the five Geotiff files were imported into Grass Gis and exported to create a csv file. Two files were created one for training and other for testing/prediction. The training data was the same regions used as A and A' and the entire region was used for testing.

### 4.2.1 Comparison of Boosted Regression Trees & Image Analogy

As shown in *Figure* 4.1, IA was carried out in two phases, one for photorealistic images and other for classified images. The B' in the former case generated a photorealstic image which was in RGB color space while the B' generated in latter case was a binary image. So, in order to compare both the B's to the originally classified image obtained from Google Earth Engine, the B' from IA of photorealistic image was further processed using the unsupervised training feature in Grass GIS which produced a binary classified image.

In order to make the results of BRTs comparable to the originally classified image obtained from Google Earth Engine, the predicted results obtained on the test data of the entire

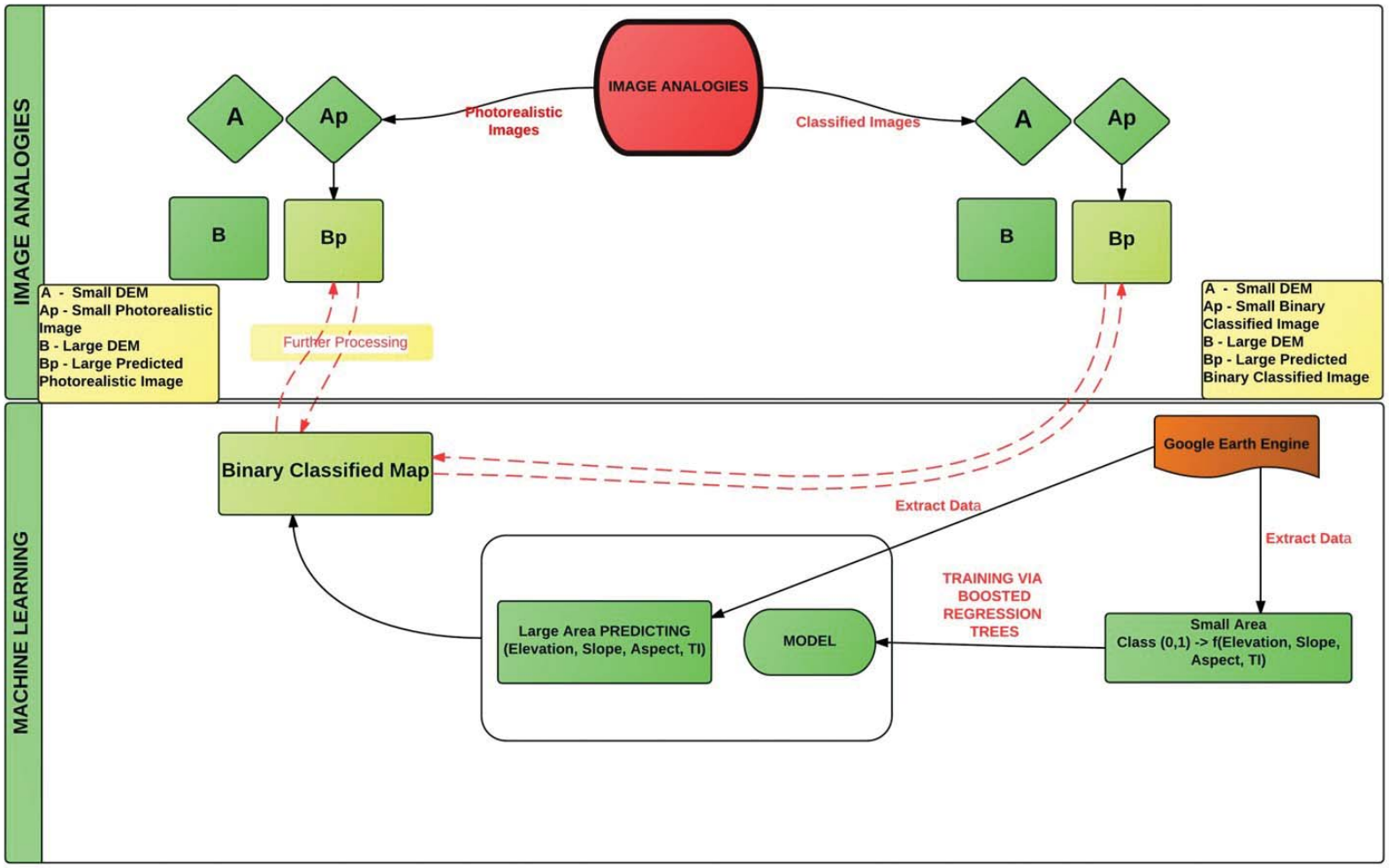



**Figure 4.1:** Comparison of Boosted Regression Trees & Image Analogy

region was in form of probability which represented the probability of presence of forest at that instance. So, 0.5 was chosen as the threshold point. An instance with a probability below 0.5 was classified as non-forest(0) and an instance with a probability greater than 0.5 was classified as forest(1). Once the predicted data was available in form of 0s and 1s, a raster data set was created using the predicted classes and imported into Grass GIS. The image obtained from the predicted class was used for comparison to the originally classified image.

# CHAPTER V

# IMAGE ANALOGY RESULTS

IA tests were carried out for various regions across continental United States to capture the varying climatic, topographical conditions. The following sections will show the results of IA.

## *5.1 Site Description*

### 5.1.1 William B. Bankhead National Forest, Alabama

The William B. Bankhead National Forest located in Northwestern Alabama lies in the Cumberland plateau region [60]. The extent of the chosen area is 32.2 km x 18.3 km and A and A' is 20 percent of the total region. Its topography consists of elevated bluffs and sloped ridges that gives rise to steep gorges, waterfalls and streams [60]. The flora primarily consists of old hardwood trees such as oak, maple, beech, and black gum and pines [60]. The annual temperature ranges from 15 °F to 95 °F with an average temperature of 65 °F [61, 62, 63].This region recorded an annual precipitation of 70 inches [64]. The elevation ranged from 169 meters to 313 meters [65].

### 5.1.2 Ochoco National Forest, Oregon

The Ochoco National Forest located in central Oregon consists of rimrock and canyons. The vegetation here ranges from dense pine forests to deserts in elevated regions [66]. The region receives an annual precipitation of about 20 to 25 inches and the average annual temperature here is about 40 °F from a minimum of -25 °F to maximum of 105 °F [64, 61, 62, 63]. The elevation ranges from 447 meters to 2397 meters [65].

### 5.1.3 Mt. Baker, Washington

Mt. Baker situated in North western Washington in Northern Cascades region consists glaciated active volcanic region. This region is dominated by hemlock and Douglas Fir tree

species [67]. The elevation ranged from 238.42 meters to 2621.5 meters [65]. The soil type in this region consists of andisols and inceptisols [68, 69]. The annual total precipitation ranges from about 30 to 50 inches [64]. The annual mean temperature was recorded at 65 °F and the range was from 5 °F to 105 °F [61, 62, 63].

#### 5.1.4 Los Alamos, New Mexico

Los Alamos is located in north central New Mexico, in the east part of the Jemez Mountains. The predominant vegetation consists of ponderosa pines and mixed conifer, spruce and fir tree types [70]. The elevation ranges from 1626 to 3030m in an area of 534 square kilometers region, including entisols, inceptisols and alfisols, there is also the presence of exposed rock formation [65, 71]. The annual average rainfall precipitation is 18.9 inches [64]. The annual mean temperature is 48.35°F, with an annual average high temperature of 59.9°F and a low temperature of 36.8°F [61, 62, 63].

#### 5.1.5 Baraboo, Wisconsin

The Baraboo region located in the central eastern portion of Wisconsin has an expanse of 184 km x 104 km. This region consists of Proterozoic-aged Baraboo quartzite rising in the form a doubly plunging syncline [72]. The Baraboo Hills were formed due to glacial action and it has a very contrasting topography [72]. It is one the largest habitats of hardwoods and consists of a mixture of conifer-deciduous trees [73]. The annual maximum and minimum temperatures here are 95 °F and -5 °F respectively and an annual average of 45 °F with a total precipitation of 40 inches [61, 62, 63, 64]. Andisols are the dominant soil types in Baraboo region [71].

#### 5.1.6 Prescott, Arizona

Prescott is located in Bradshaw mountains in north central part of Arizona and this study region has an expanse of 85 km by 46.5 km [74]. The elevation in this region ranges from 901 meters to 2235 meters [65]. The annual precipitation is about 20 to 25 inches and the annual average temperature is 65 °F [64, 61].The minimum and maximum annual temperature ranges from 5 °F and 105 °F [62, 63]. Alfisols and vertisols are major dominating soil types

[71].

### 5.1.7 Waynesville, North Carolina

Waynesville located in North Carolina is located close to the Great Smoky Mountains National Park and the Blue Ridge Parkway. This region is located in a valley amongst peaks [75]. This region is has an extent of 170 km by 93 km and the elevation ranges from 289 meters to 1886 meters [65]. Histosols and ultisols are the dominant types of soils in this region [71]. The total precipitation is 70 inches [64]. The annual mean temperature is 55 °F, with an annual high temperature of 95 °F and a low temperature of 15 °F [61, 62, 63].

### 5.1.8 Reno, Nevada

Reno is situated in the northwestern part of Nevada and the extent of this study region is 170 km by 93 km. Wetlands are an essential part of this region. Reno lies on the western edge of Great Basin and the rain shadow side of Sierra Nevada with numerous faults interspersing this region [76]. The average annual temperature here is about 45 °F with a minimum and maximum of -15 °F and 105 °F [61, 62, 63]. The total precipitation recorded is about 6 inches [64]. Ardisols, alfisols and entisols are the major soil types present here [71].

### 5.1.9 Cincinnati, Ohio

Cincinnati is located on the border of Kentucky and Ohio where the Ohio and Licking river meet. This study region has a massive extent of 380 km by 184 km. This region located is the Ohio river valley primarily consists of hills, bluffs and low ridges [77]. The elevation in this region ranges from 145 meters to 346 meters [65]. This region has an annual average temperature of about 45 °F with a total precipitation of 45 inches [61, 64]. The maximum and minimum temperatures were recorded at 95 °F and 5 °F [63, 62]. The type of soil present here is alfisols [71].

### 5.1.10 Aspen, Colorado

Aspen, Colorado is located in between the Rocky and Elk Mountains [78]. Aspen, pine, spruce and fir trees are dominant forms of vegetation here [79]. This study region spreads over an area of 184 by 104 square kilometers with an elevation range of 1820 meters to 4220

meters [65]. This region has an annual average temperature of 45 °F with a minimum and maximum of -35 °F and 95 °F [61, 63, 62]. This region receives a total rainfall of about 20 inches [64]. This region is dominated by mollisols type of soil [71].

#### 5.1.11 Mt. Linn, California

Mount Linn is located in the Northern Coastal ranges of California. This area has an extent of 195 km by 105 km and the elevation ranges from 206 meters to 2340 meters [65]. The major species in this region are juniper, pine, hemlock, fir, douglas and cedar [80]. The annual average temperature of the region is 65 °F with a maximum of 115 °F and a minimum of 25 °F [61, 63, 62]. This region receives a total rainfall of 10 inches [64]. Histosols are the dominant soil types in this region [71].

#### 5.1.12 Rio Grande Forest, Colorado

Rio Grande Forest is located in southwestern Colorado and contains both agricultural alpine as well as high deserts [81]. This area has an extent of 195 km by 105 km and the elevation ranges from 2127 meters to 4152 meters [65]. The annual average temperature of the region is 45 °F with a maximum of 85 °F and a minimum of -25 °F [61, 63, 62]. This region receives a total rainfall of 20 to 25 inches [64]. Alfisols and mollisols are the dominant soil types in this region [71].

#### 5.1.13 Hammersley Wild Area, Pennsylvania

Hammersley situated in north central Pennsylvania majorly consists of second growth forests. The vegetation growth here mainly consists of hemlocks and pines [82] . This area which looks mountainous is actually made up of eroded tops to form a dissected plateau. The elevation ranged from 164.067 to 759.1491 meters [65]. The soil type in this region consists of well drained inceptisols [69]. The annual total precipitation ranges from about 30 to 40 inches [64]. The annual mean temperature was recorded at 45 °F and the range was from -5 °F to 95 °F [61, 62, 63].

## *5.2 Qualitative Analysis*

### 5.2.1 William B. Bankhead National Forest, Alabama

*Figure* 5.1 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

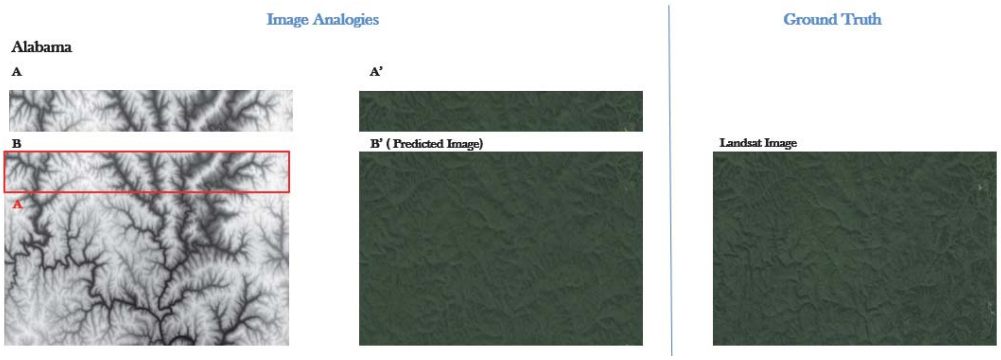



**Figure 5.1:** Image Analogy of Alabama

As seen in *Figure* 5.1, the region of Waynesville is originally seen to be densely forested and it has been adequately captured by the IA in B'. On a closer look, it can analyzed that B' has also successfully captured the topographical texture of the region as seen in the form of ridges in A'.

### 5.2.2 Ochoco National Forest, Oregon

*Figure* 5.2 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

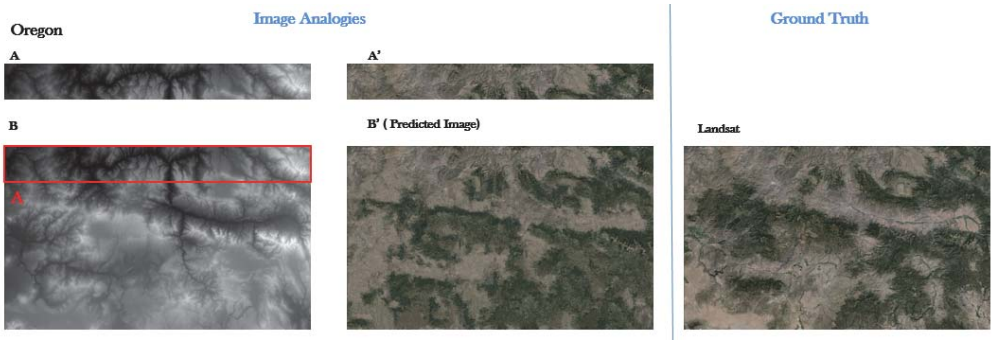



**Figure 5.2:** Image Analogy of Oregon

As seen in *Figure* 5.2, the Ochoco site consists of an arid region with vegetation patches in areas of higher elevation. The IA reproduces the biomass in the predicted image(B') and also successfully preserves the texture of the original Landsat image in B'. On comparison of B' with the actual image, it is apparent that the technique creates extra biomass than that is present originally. Here again, the SSIM value obtained is 0.301 which indicates a low match contrary to the visual observations.

### 5.2.3 Mt. Baker, Washington

*Figure* 5.3 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

At the Mt. Baker(WA) site, on comparison of predicted image with Landsat image, we can determine from *Figure* 5.3 that the IA does a good job of creating a visual image from the Digital Elevation Map. The forested upland region which forms a major part of the site has been appropriately captured in B'. Also, the predicted image captures most of the river that passes through the site with very few points of error. Not only has the IA captured forested areas but also the dry regions that are areas of lower elevation which lie on the eastern part of the site. Although the demarcation between ridges and valleys seen on the

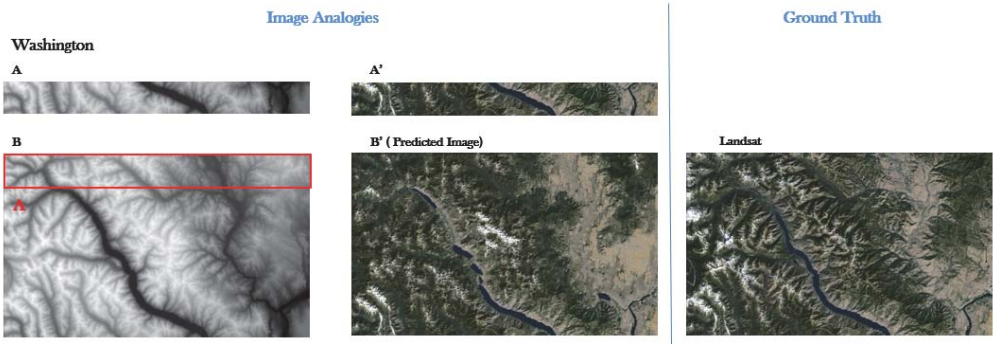



**Figure 5.3:** Image Analogy of Washington

western side of region on Landsat image has not been captured very well on B'.

### 5.2.4 Los Alamos, New Mexico

*Figure* 5.4 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

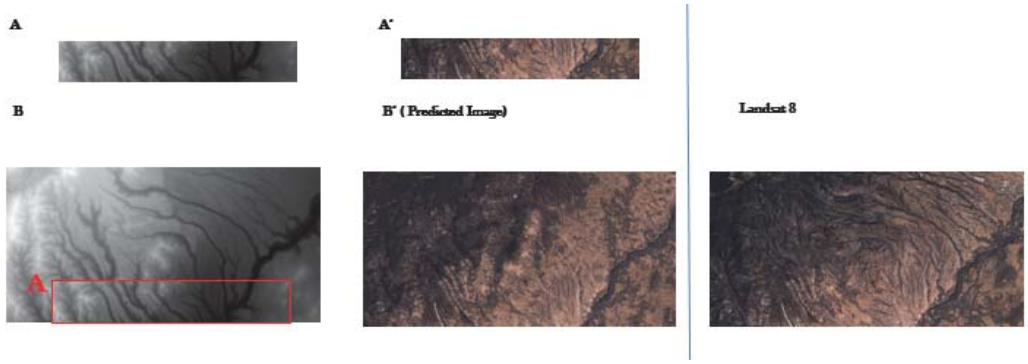



**Figure 5.4:** Image Analogy of New Mexico

The Los Alamos region again is a primarily dry area with vegetation patches on the north western corner which is at higher elevation Figure 5.4. The IA successfully recreates the

vegetation cover in predicted image B' as in the original image. Also, the drier regions which dominates the region has been reproduced in B'. Though the IA fails at preserving the texture of the original image as the dried up streams seen in the actual image are not recreated in B'.

### 5.2.5 Baraboo, Wisconsin

*Figure* 5.5 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

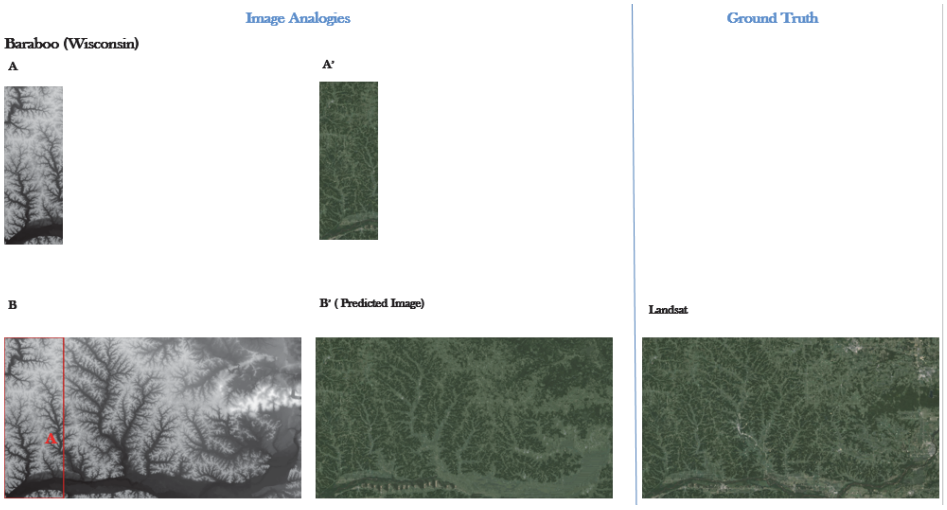



**Figure 5.5:** Image Analogy of Wisconsin

As seen in *Figure* 5.5, the region in Wisconsin consists of densely forested regions in the valleys and grasslands in the uplands. This test was done in order to see how well the IA works in an inhabited area. As evident from the result of B', it is a very good representation of the original Landsat Image minus the urbanized colonies. B' has slightly increased forested areas in the upper right corner and if seen closely the dense forest region

along the deeper parts of valley in bottom right corner has been wiped out too.

### 5.2.6 Prescott, Arizona

*Figure* 5.6 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

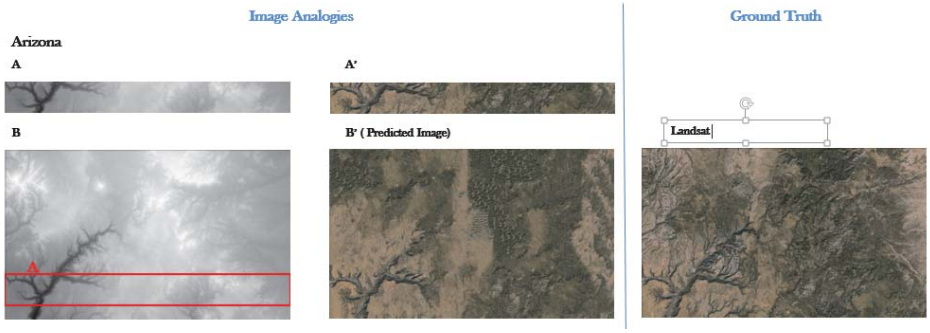



**Figure 5.6:** Image Analogy of Arizona

The region of Prescott is mostly a plain land with the exception of one ridge in the its south-west corner. Also, this region is mostly dry and has only sparse vegetation. This case is an example of failed IA as the B' lacks texture and only the sample areas used as training pairs have been correctly reproduced.

### 5.2.7 Waynesville, North Carolina

*Figure* 5.7 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

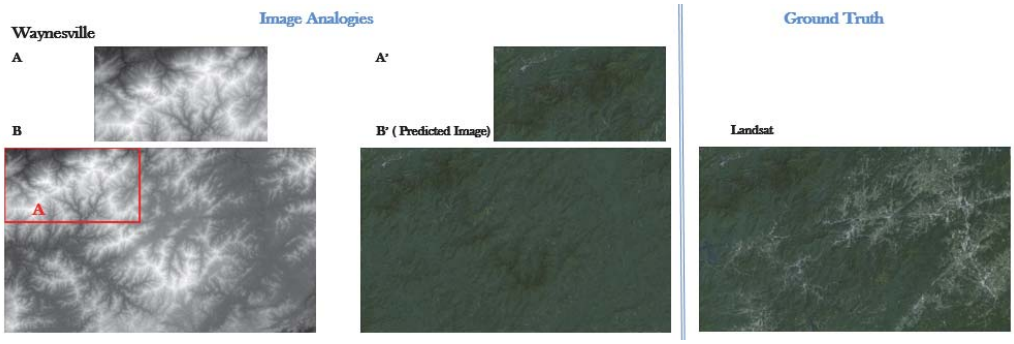



**Figure 5.7:** Image Analogy of North Carolina

Like Baraboo, Waynesville again was a test of IA in urbanized areas. *Figure* 5.7 shows that this region is composed of valleys and uplands and it is mostly forested. The forest is otherwise interspersed by inhabited regions in the east. Here again, the IA does correctly capture the vegetation in the original Landsat image and the urbanized areas have been replaced by forests in B'.

### 5.2.8 Reno, Nevada

*Figure* 5.8 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

This region is Reno is forested on the western side which is slightly at a higher elevation compared to the rest of the region. Although the western region of the original image has been correctly captured in B', but the eastern region which was dry and barren has been over forested in B'.

### 5.2.9 Cincinnati, Ohio

*Figure* 5.9 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part.

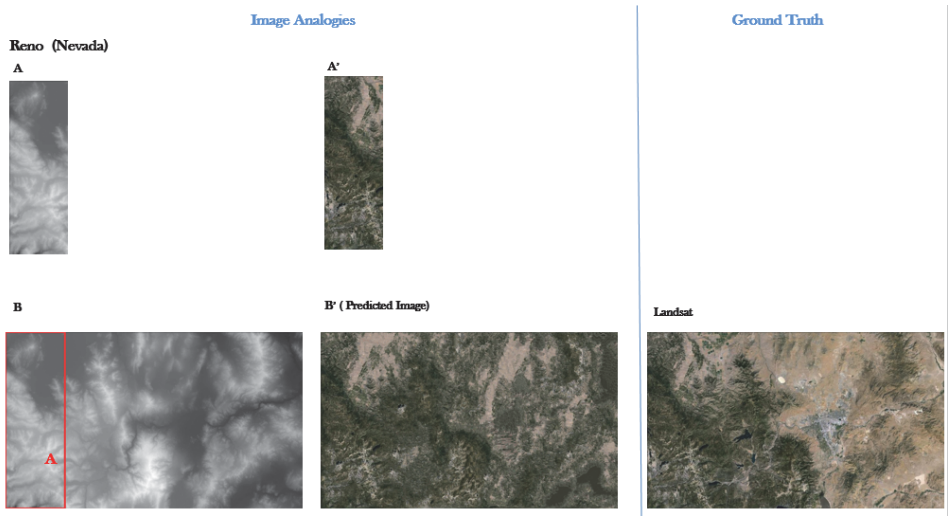



**Figure 5.8:** Image Analogy of Nevada

The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

This region in Cincinnati has a deep ridge running east-west through the center of the region and has an urbanization in the north-west corner. It is densely forested in the eastern side and it gives way to grasslands in the central region. Although as seen from *Figure* 5.9, B' retains the texture of the whole region but because the sample area, A' has mostly forested region, B' too is over vegetated.

### 5.2.10 Aspen, Colorado

*Figure* 5.10 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

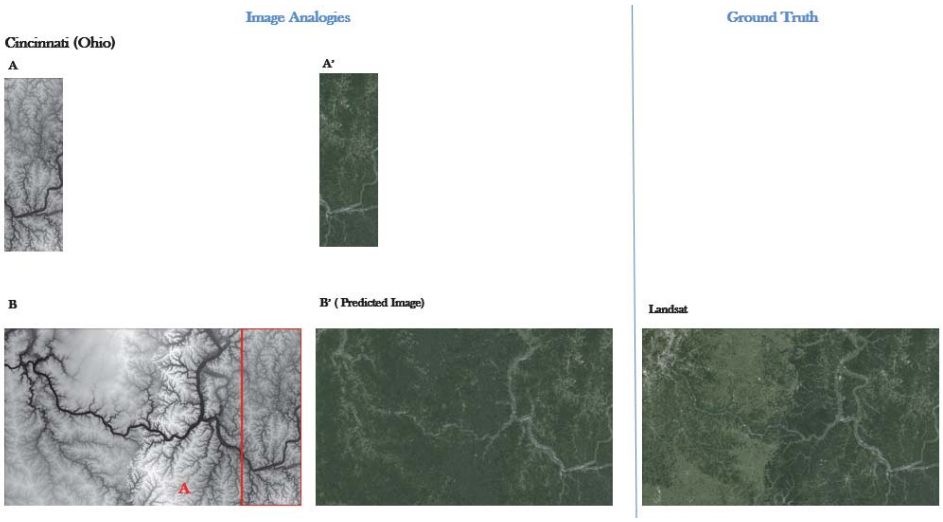



**Figure 5.9:** Image Analogy of Ohio

The region of Aspen, Colorado consists of barren uplands and vegetated valleys. The image B' is a good representation of both texture and vegetation in the original Landsat image.

### 5.2.11 Mt. Linn, California

*Figure* 5.11 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

Mt. Linn, California is overall a forested region except the dry patch in the east running north-south. The IA in this region as seen from *Figure* 5.11 has a good performance. B' is a very close representation of the original Landsat image, except towards the center of the image, the B' has a bit too much vegetation.


**Figure 5.10:** Image Analogy of Colorado

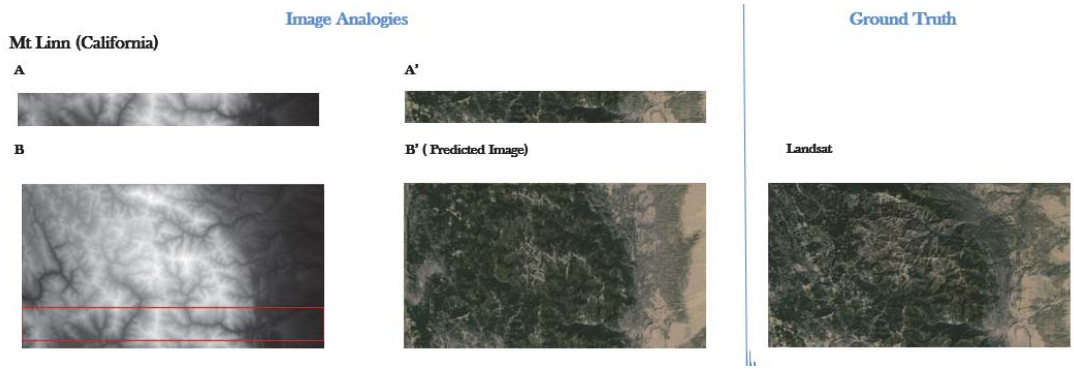



**Figure 5.11:** Image Analogy of California

### 5.2.12 Rio Grande Forest, Colorado

*Figure* 5.12 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

This region in Rio Grande Forest, Colorado is again a dry region with patches of vegetation. However, the IA in this region as seen from *Figure* 5.12 is a case of failure as B' has just created the dry and green patches but neither is it in right amount and the texture is


**Figure 5.12:** Image Analogy of Colorado

also missing in comparison to the Landsat image.

### 5.2.13 Hammersley Wild Area, Pennsylvania

*Figure* 5.13 is a representation of the IA of this region where A (DEM) and A'(photorealistic image) is the training pair of images and B is the larger input DEM of which A is a part. The image B' was synthesized from B using the relation between A and A'. The ground truth is the original Landsat image obtained from Google Earth Engine for comparing the generated B'.

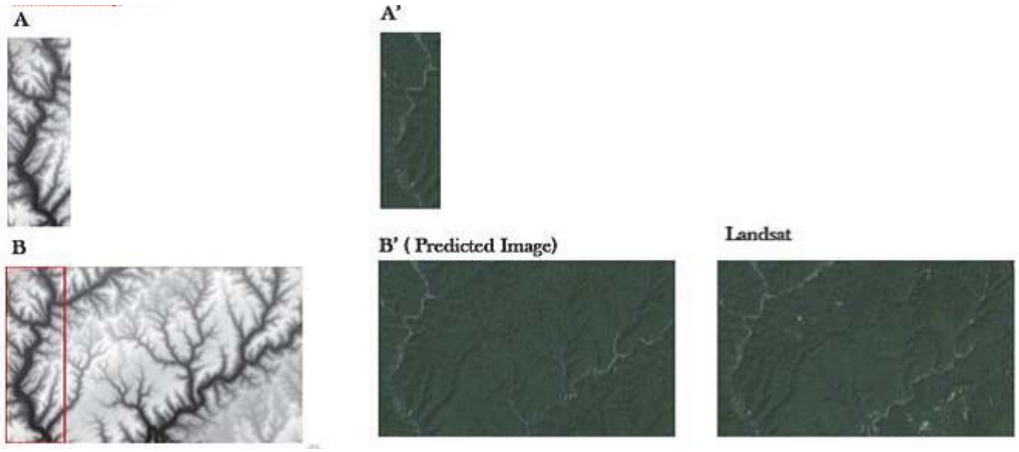



**Figure 5.13:** Image Analogy of Hammersley

The Hammersley site as seen in Figure 5.13 is a dense forested area lying in the upland region. On comparison the predicted image B' with the original Landsat image, it can be seen that the IA does a good job at predicting vegetation cover of the region from the its Digital Elevation Map (DEM). Also, the streams seen as darker regions in DEM(B) have been well captured. The south eastern part of Landsat shows small patches of dry uplands which has been lost in B'. The overall texture of the actual image has been very well preserved in B'.

## *5.3 Quantitative Analysis*

The image quality of all the predicted images for the regions was measured using various IQA metrics and the results for the same have been tabulated below in *Table* 5.1. For each of the IQA metrics a higher value indicates higher image quality except for MSE where higher value indicates more distortion. *Table* 5.2 shows the ranking of predicted image quality of the four sites using the four comparison techniques; none of the two metrics agree with each other on the quality. Thus, IA method lacks a standard of measuring the image quality.

All the methods of comparison almost unanimously show that the Hammersley(PA) and William Bankhead Forest(AL) regions have the best predicted image in comparison to its original image and this result in congruity to the one's visual observation as to which predicted image is the most photo-realistic image when compared to the actual image.

**Table 5.1:** Values of IQA Metrics at different sites

| SITE | MSE | UIQI | SSIM | FSIM |
|---|---|---|---|---|
| William B.(AL) | 18.905429 | 0.996167 | 0.781639 | 0.882472 |
| Ochoco(OR) | 93.83793 | 0.939619 | 0.354246 | 0.764015 |
| Mt. Baker(WA) | 86.39245 | 0.725614 | 0.078765 | 0.62138 |
| Los Alamos(NM) | 79.25642 | 0.9481 | 0.29768 | 0.728792 |
| Baraboo (WI) | 67.163652 | 0.966195 | 0.506994 | 0.793139 |
| Prescott(AZ) | 85.585659 | 0.955251 | 0.298308 | 0.740355 |
| Waynesville (NC) | 68.144466 | 0.972578 | 0.612453 | 0.720364 |
| Reno(NV) | 89.816533 | 0.918814 | 0.32882 | 0.731117 |
| Cincinnati (OH) | 99.016418 | 0.960463 | 0.483491 | 0.783014 |
| Aspen (CO) | 81.175672 | 0.929363 | 0.344963 | 0.738776 |
| Mt. Linn(CA) | 82.694889 | 0.910458 | 0.31211 | 0.731822 |
| Rio Grande(CO) | 110.83028 | 0.925364 | 0.26342 | 0.71761 |
| Hammersley(PA) | 12.60859 | 0.9951 | 0.813693 | 0.8374 |

**Table 5.2:** Ranking of Sites based on different IQA metric values

| SITE | MSE | UIQI | SSIM | FSIM |
|---|---|---|---|---|
| William B.(AL) | 2 | 1 | 2 | 1 |
| Ochoco(OR) | 11 | 8 | 6 | 5 |
| Mt. Baker(WA) | 9 | 13 | 13 | 13 |
| Los Alamos(NM) | 5 | 7 | 11 | 10 |
| Baraboo (WI) | 3 | 4 | 4 | 3 |
| Prescott(AZ) | 8 | 6 | 10 | 6 |
| Waynesville (NC) | 4 | 11 | 8 | 9 |
| Reno(NV) | 10 | 11 | 8 | 9 |
| Cincinnati (OH) | 12 | 5 | 5 | 4 |
| Aspen (CO) | 6 | 9 | 7 | 7 |
| Mt. Linn(CA) | 7 | 12 | 9 | 8 |
| Rio Grande(CO) | 13 | 10 | 12 | 12 |
| Hammersley(PA) | 1 | 2 | 1 | 2 |

# CHAPTER VI

# BOOSTED REGRESSION TREES

IA doesn't have one universal standard of measuring the image quality and as seen from image quality analysis, the best image varies with human perception and the various image quality metrics. Image analogy was used as a basis to decide the region to be used as training and testing data sets. Also, not all regions gave promising results which implies predicted images(B') which resembled the original landsat image. Four regions with their B' bearing close resemblance to their respective original landsat images were chosen for further analysis using Boosted Regression Trees. The selection of these regions were also on a based on the diversity in topographical, climatic and geographical conditions. Unlike IA, BRT also returned the order of dominance of the attributes. The regions chosen to carry out in depth analysis using BRT are : Los Alamos(New Mexico), Mt. Baker (Washington), Ochoco National Forest (Oregon), Hammersley (Pennsylvania).

The BRT model produces a deviance vs. number of trees fitted graph. The green vertical line in the graph depicts the number of trees in the BRT model at which maximum reduction is deviance was achieved which is marked by the red horizontal line. Another set of graphs is the Dominance of attributes which depicts how influential each attribute is. It also gives the kind of relationship, for instance, linear or non-linear, between each attribute and class.

## *6.1 Los Alamos(New Mexico)*

The New Mexico region had 511335 data points on the prediction data set of which 158752 data points were used as training data set develop a Boosted Regression Tree model. *Figure* 6.1 shows the fitted BRT model had a total of 6600 trees to reach a minimum deviance of 0.861. The model had a training accuracy of 87.3% and on predicting the land cover class over the testing data set using this model, 73% of the total instances were

correctly predicted. *Figure* 6.2 represents the order of dominance of the four attributes which are in the order of elevation(82.7%), aspect (7.9%), slope(7.3%) and topographic index (2.1%). Elevation is the most influential attribute in this region with a maximum influence at an elevation of 1500 to 2000 meters above sea level.

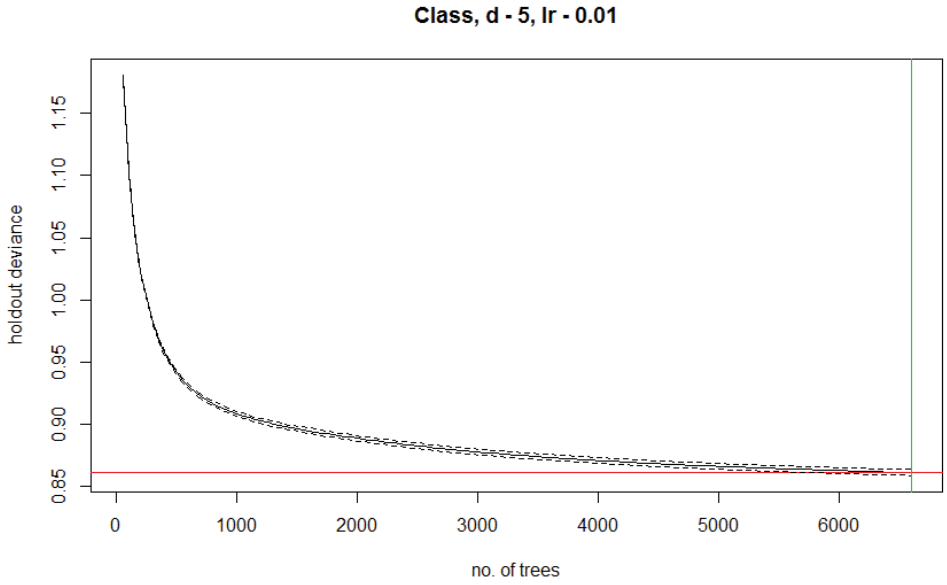



**Figure 6.1:** Trees fitted vs. Deviance for New Mexico Region

## *6.2 Mt. Baker (Washington)*

The Washington region had 881325 data points on the prediction data set of which 168298 data points were used as training data set develop a Boosted Regression Tree model. *Figure* 6.3 shows the fitted BRT model had a total of 7550 trees to reach a minimum deviance of 0.661. The model had a training accuracy of 87.3% and on predicting the land cover class over the testing data set using this model, 74.93% of the total instances were correctly predicted. *Figure* 6.4 represents the dominance of the four attributes which are in the order of aspect (46.4%), elevation(37.8%), slope(11.4%) and TI (4.3%). Here, again aspect and elevation were the two most influential attributes affecting vegetation cover. The

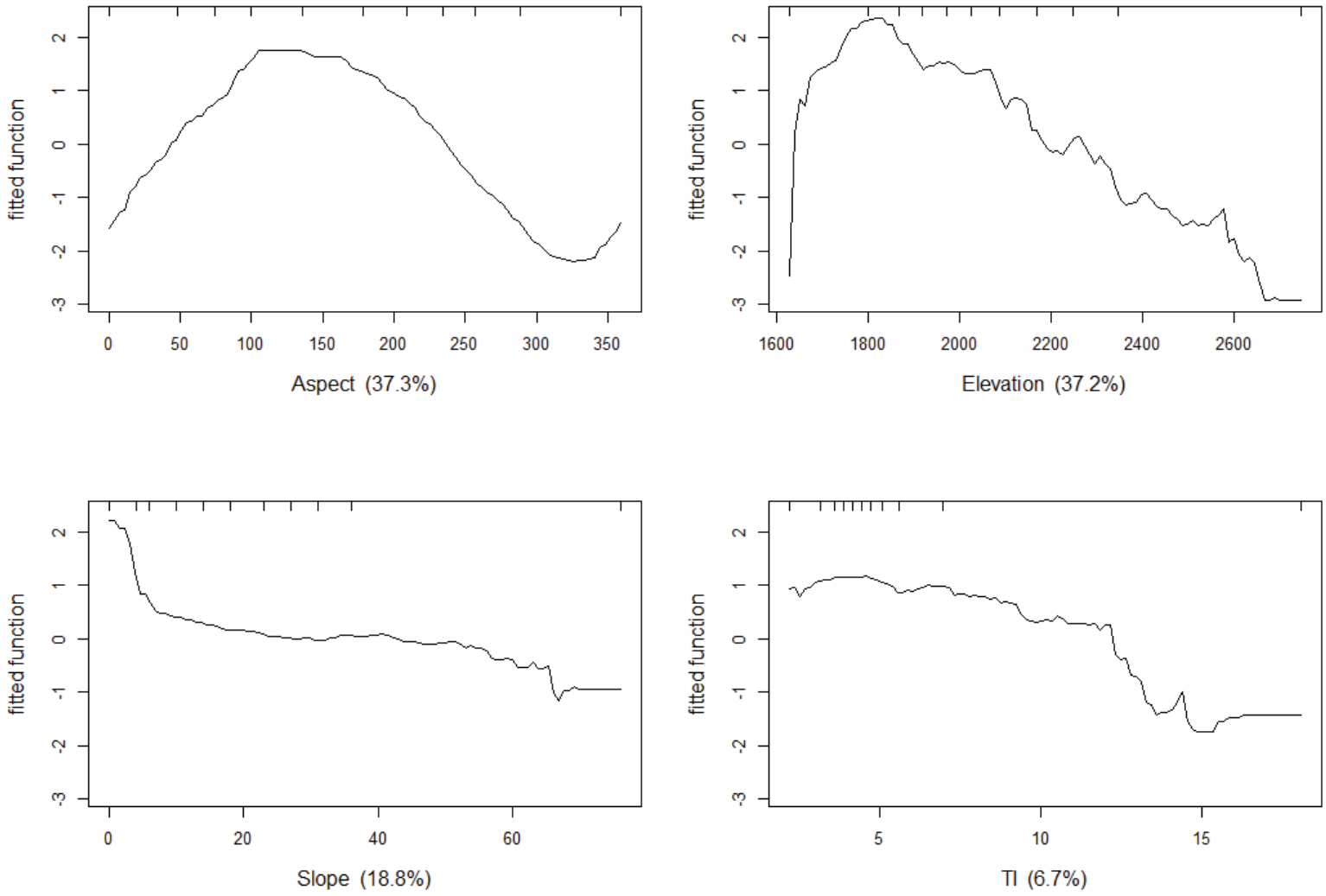



**Figure 6.2:** Dominance of attributes for New Mexico Region

slopes facing west and south-west were more prone to being forested. Within Washington region, points with the effect of elevation was recorded high for heights of 500 to 1000 meters and 2000 to 2500 meters, between which its effect subsided.

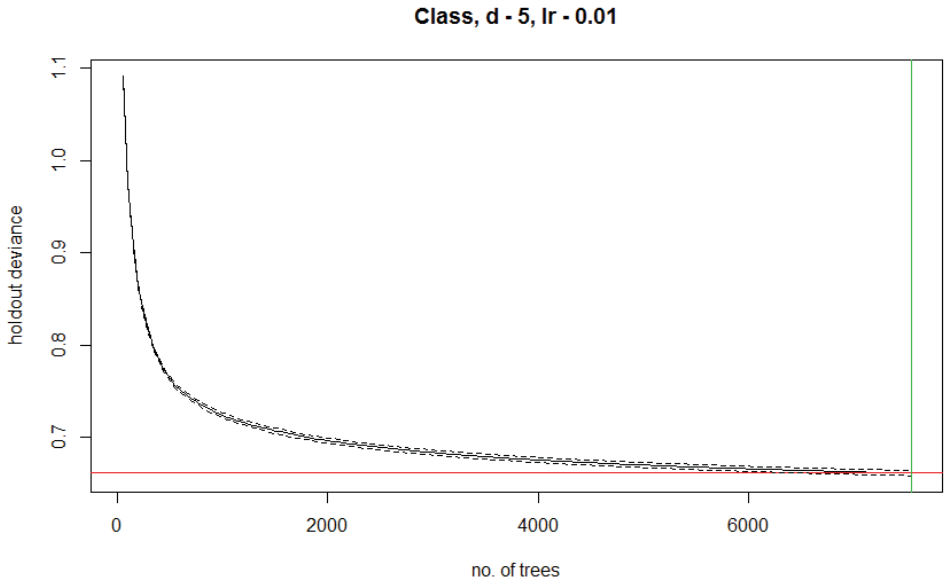



**Figure 6.3:** Trees fitted vs. Deviance for Washington Region

## *6.3 Ochoco National Forest (Oregon)*

This region located in Oregon had a test set with 927675 attributes of which the training dataset had a total of 194373 attributes, using which the BRT model was developed. As seen in *Figure* 6.5, the BRT model reduced the deviance upto 0.239 by fitting a 3150 trees. The model had a training accuracy of 98.4% and on using it for prediction over the test data set, 92.86 % of the classes were correctly predicted. *Figure* 6.6 represents the dominance of the four attributes which are in the order of aspect (51.2%), slope(42.1%), elevation(6.3%) and TI (0.4%). Here, aspect and slope were the two most influential attributes affecting vegetation cover as this region did not really any major elevation differences. The slopes facing north, north-east and east were more prone to being forested. Within this region,

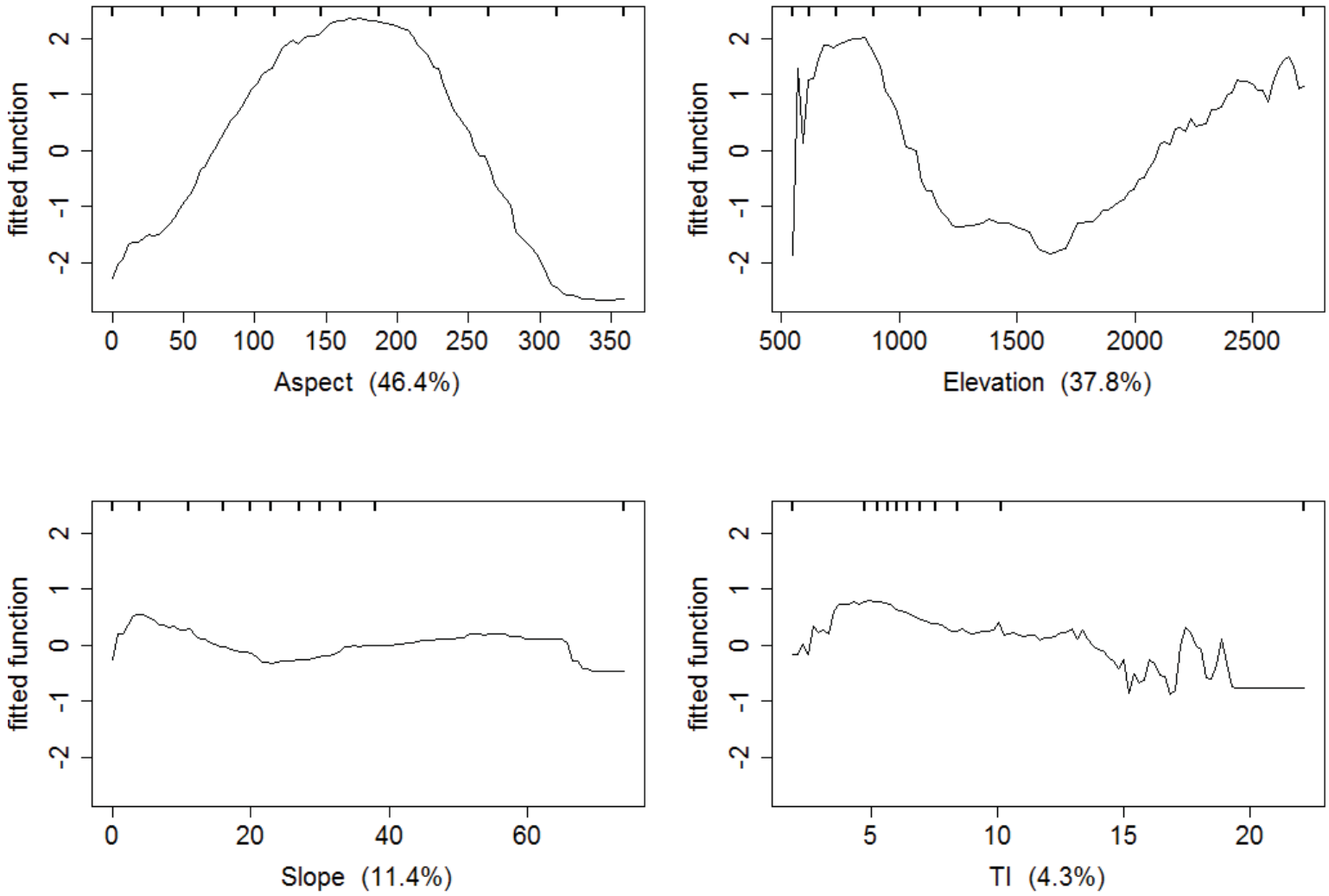



**Figure 6.4:** Dominance of attributes for Washington Region

chances of vegetation being present increased gradually with increasing slope till 30 degrees after which it became stagnant up to 50 degrees.

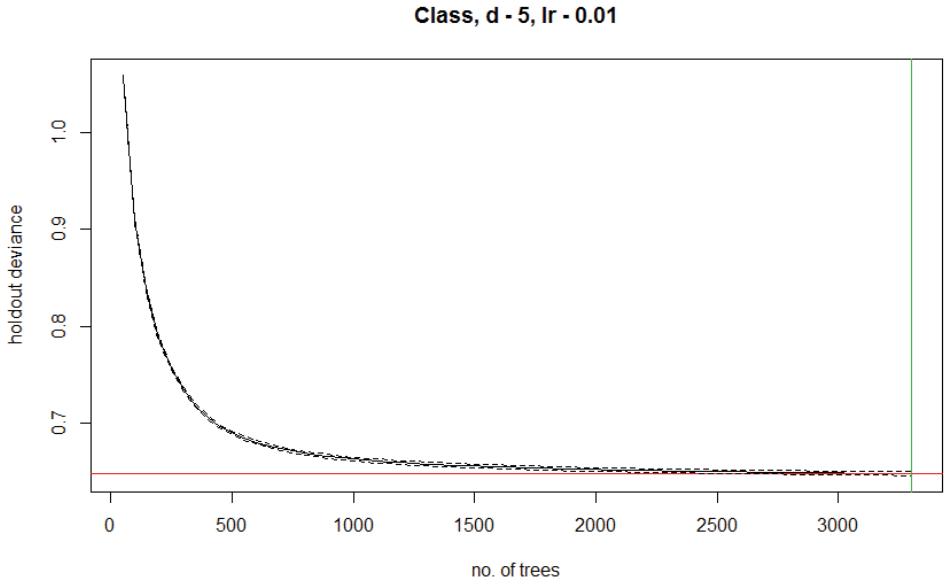



**Figure 6.5:** Trees fitted vs. Deviance for Oregon Region

## *6.4 Hammersley (Pennsylvania)*

The Pennsylvania region had a test set with 2805300 attributes of which the training dataset had a total of 436456 attributes, using which the BRT model was developed. As seen in *Figure* 6.7, the BRT model reduced the deviance upto 0.239 by fitting a 3150 trees. The model had a training accuracy of 98.4% and on using it for prediction over the test data set, 92.86 % of the classes were correctly predicted. *Figure* 6.8 represents the dominance of the four attributes which are in the order of aspect (51.2%), slope(42.1%), elevation(6.3%) and TI (0.4%). Here, aspect and slope were the two most influential attributes affecting vegetation cover as this region did not really any major elevation differences. The slopes facing north, north-east and east were more prone to being forested. Within this region, chances of vegetation being present increased gradually with increasing slope until 30 degrees

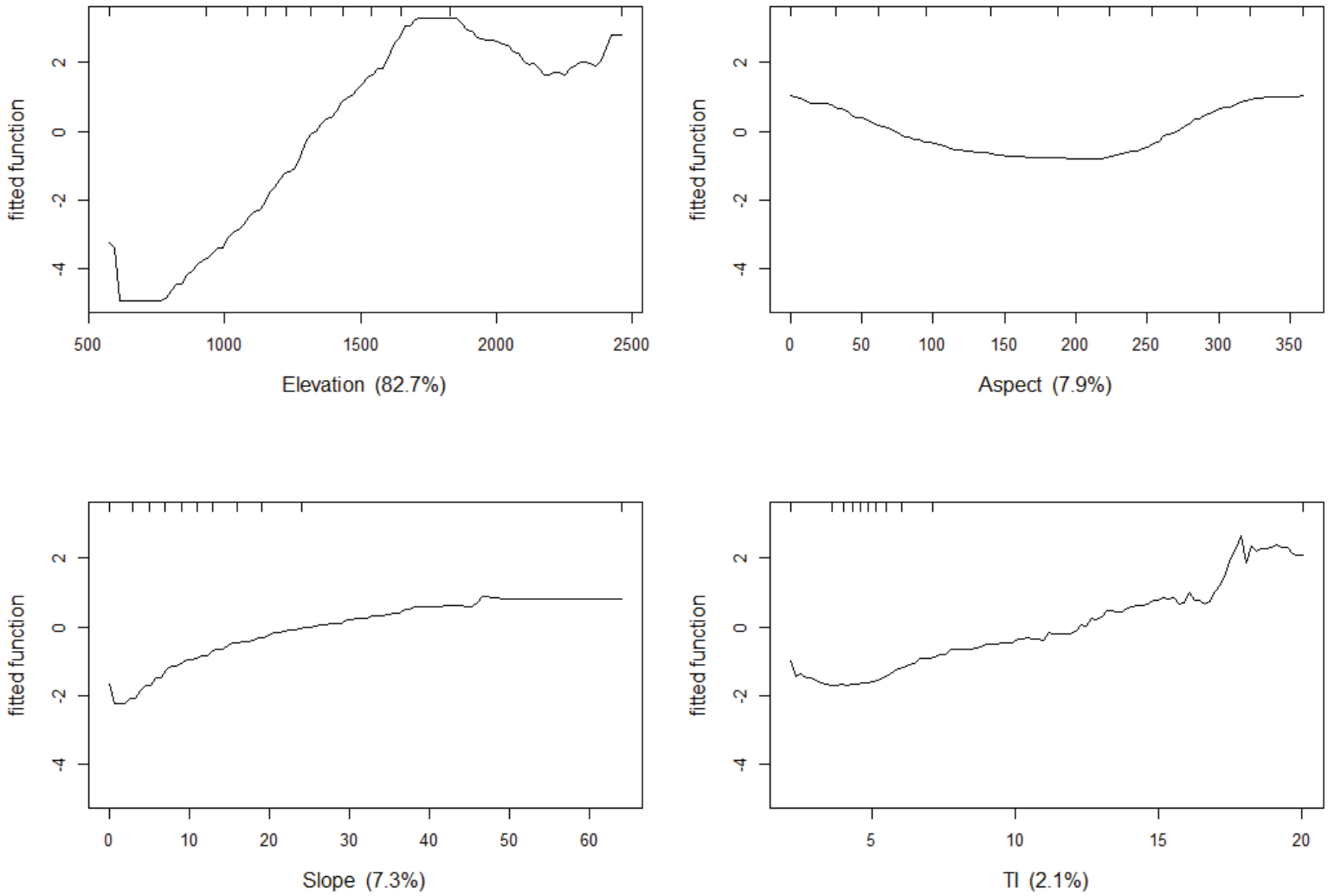



**Figure 6.6:** Dominance of attributes for Oregon Region

after which it became stagnant up to 50 degrees.

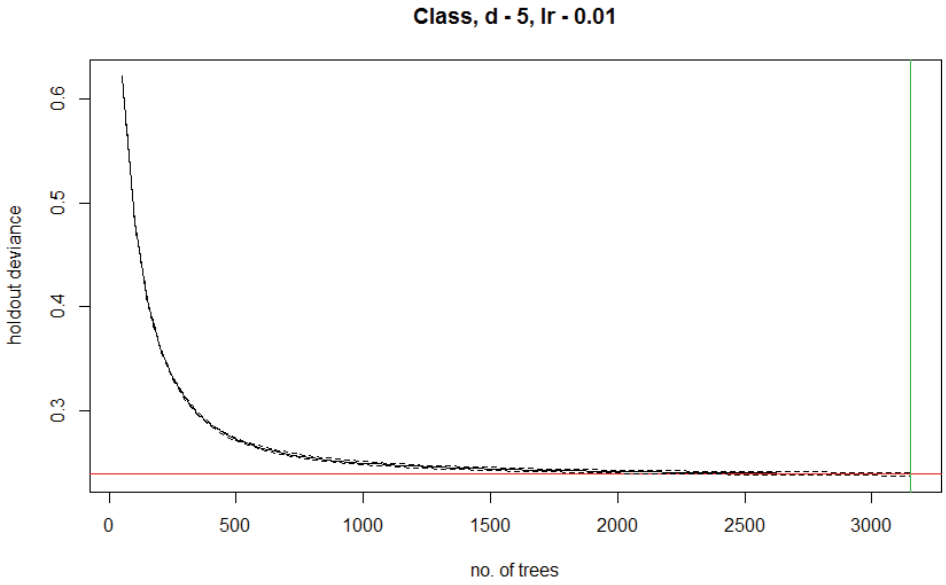



**Figure 6.7:** Trees fitted vs. Deviance for Pennsylvania Region

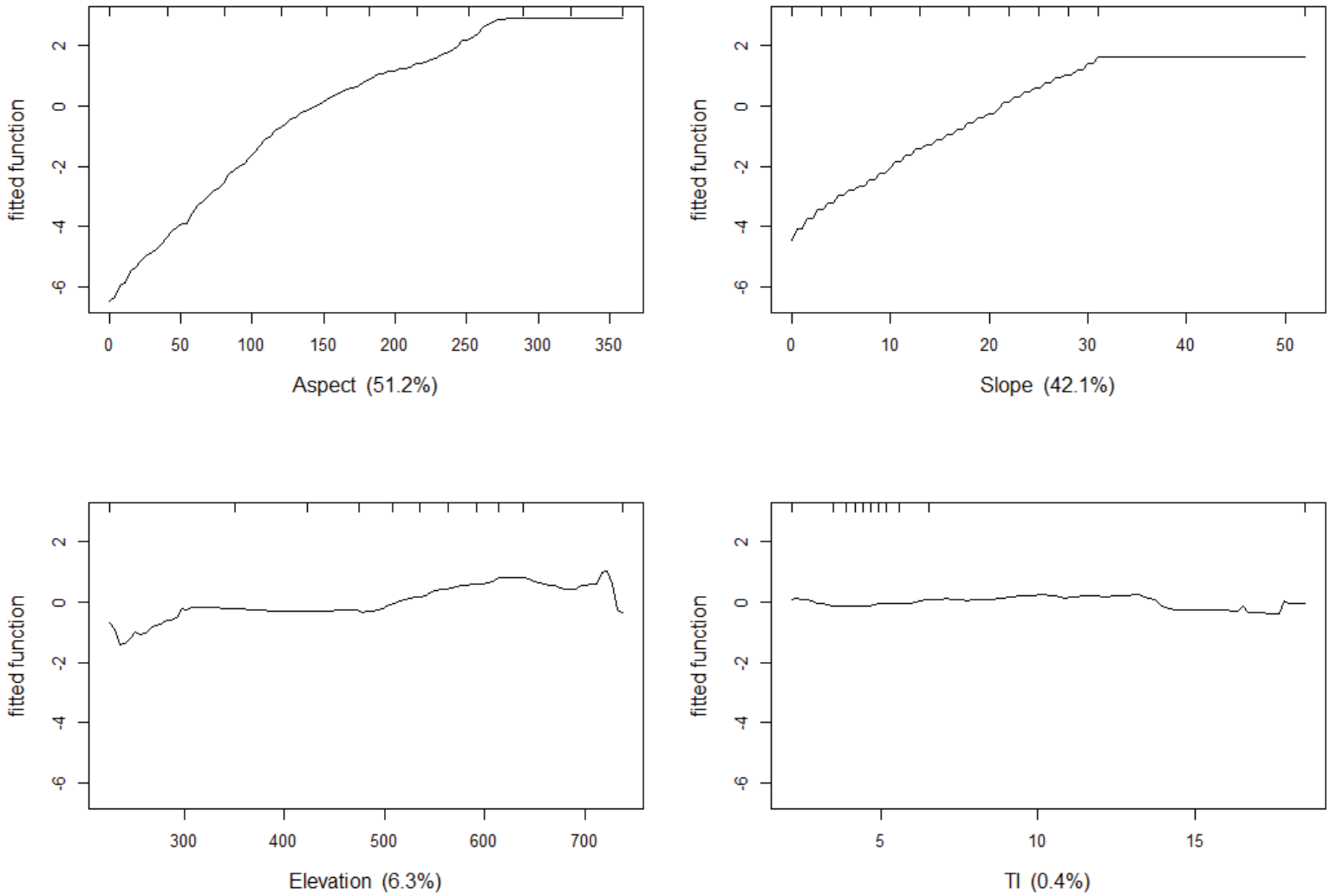



**Figure 6.8:** Dominance of attributes for Pennsylvania Region

# CHAPTER VII

# COMPARISON OF IMAGE ANALOGY & BOOSTED REGRESSION TREES

The objective of this chapter is to compare the two ML approaches that has been used i.e.; IA and BRT and to find out the best method for prediction of vegetation cover on a large scale. Image analogy itself was carried out in two ways; first using digital elevation maps and landsat images and second using digital elevation maps and classified images. As already described in *Section* 4.2.1, the images from three techniques has to be brought to one common ground before a comparison could be established. Here, all the images were converted into a binary classified image, so that it could be compared with the originally classified image. The cases for the four regions have been presented in the following sections.

## *7.1 Los Alamos(New Mexico)*

*Figure* 7.1 shows the comparison of binary classified images using all three techniques. On a pixel to pixel comparison of the originally classified Google Earth Engine image to the predicted image generated using the ML techniques, various levels of accuracies were obtained. BRTs with binary classification had the maximum accuracy of 73%, followed by the predicted image using IA using classified image with an accuracy of 66% and IA using landsat image returned the least accuracy of 53%.

## *7.2 Mt. Baker (Washington)*

*Figure* 7.2 shows the comparison of binary classified images obtained from IA and BRT. Pixel to pixel comparison of each of the images obtained using ML to the originally classified image obtained from Google Earth Engine was carried out to give the prediction accuracy of each technique. BRT returned the highest prediction accuracy of 74.94 %, followed by IA using classified image with an accuracy of 70.64 % and IA using landsat image returned

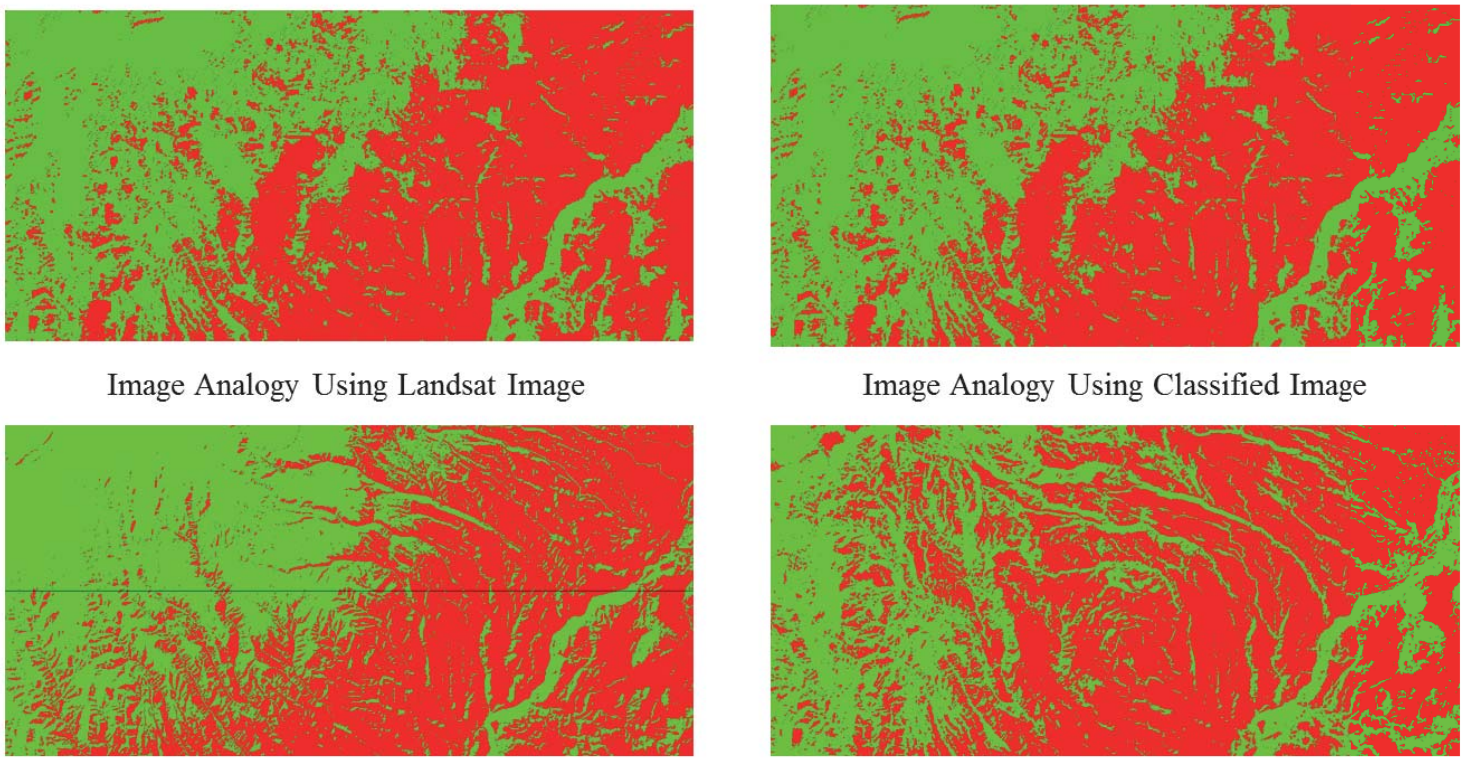



**Figure 7.1:** Comparison of Binary Classified Images for New Mexico Region

the least accuracy of 55.40 %.

## *7.3 Ochoco National Forest (Oregon)*

As shown in *Figure* 7.3, on pixel to pixel comparisons of the various predicted images obtained through IA and BRTs to the originally classified image, varying accuracies were obtained. The accuracies obtained are as follows; 81.15 %, 78.20 % and 75.59 % for BRT, IA using classified image and IA using landsat image respectively.

## *7.4 Hammersley (Pennsylvania)*

*Figure* 7.4 shows the comparison of binary classified images obtained from IA and BRTs. Pixel to pixel comparison of each of the images obtained using ML to the originally classified image obtained from Google Earth Engine was carried out to give the prediction accuracy of each technique. BRTs returned the highest prediction accuracy of 92.86 %, followed by IA using classified image with an accuracy of 77.85 % and IA using landsat image returned the least accuracy of 48.70 %.

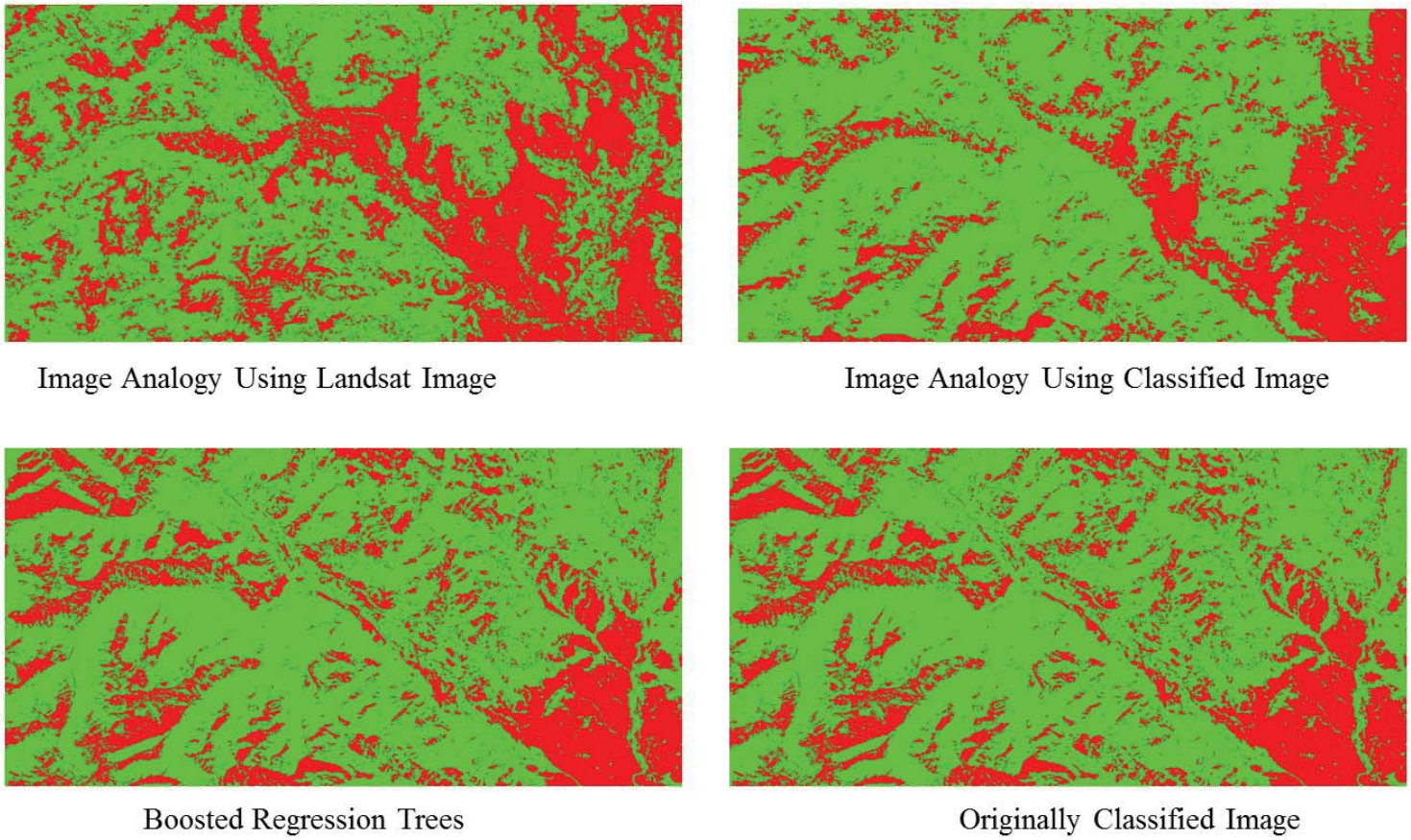



**Figure 7.2:** Comparison of Binary Classified Images for Washington Region

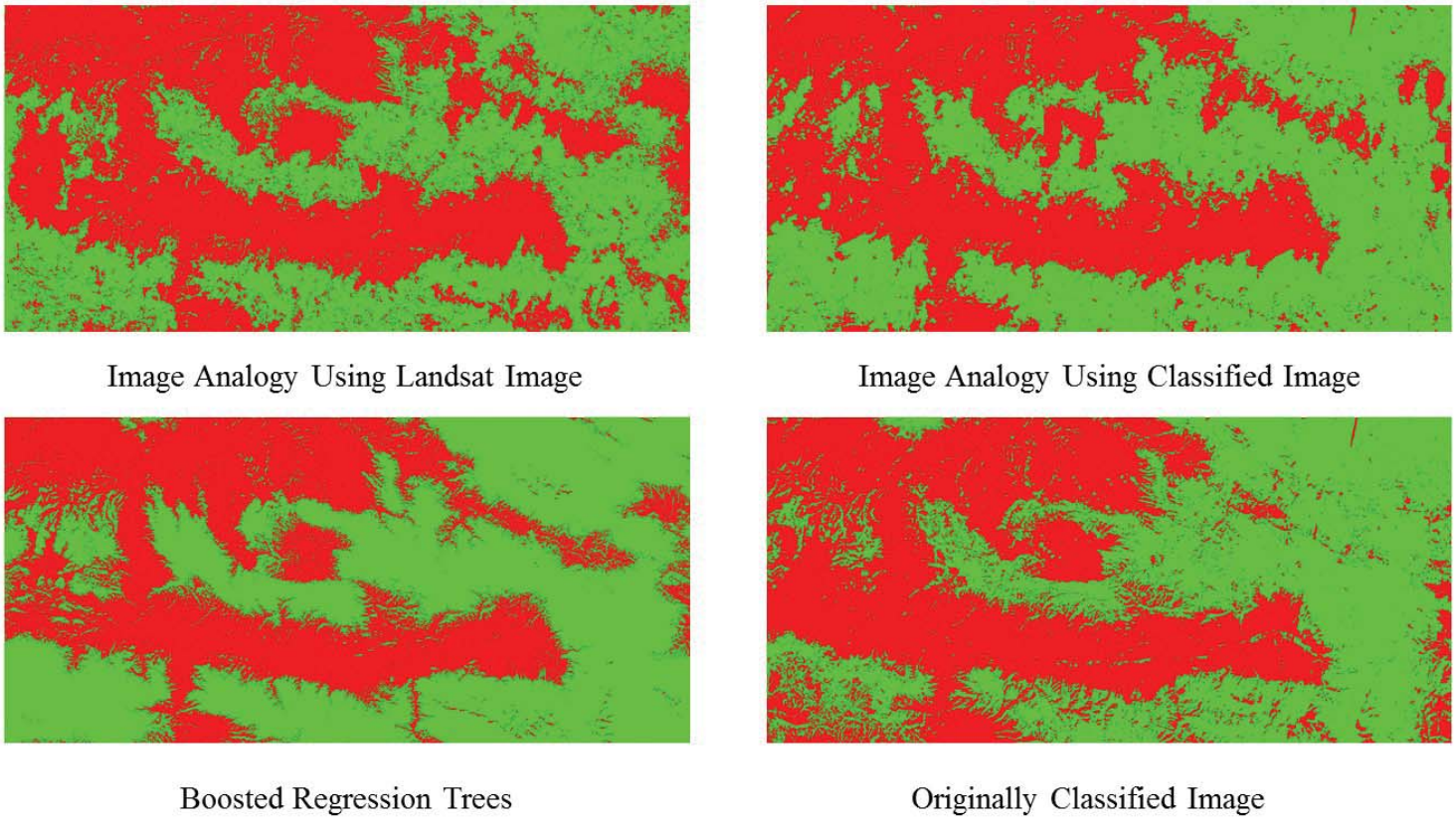



**Figure 7.3:** Comparison of Binary Classified Images for Oregon Region

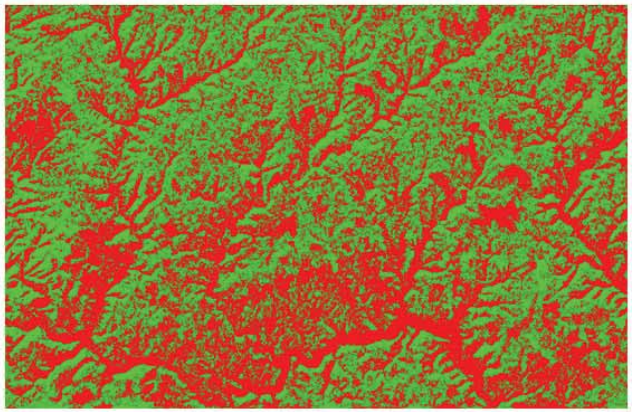


Image Analogy Using Landsat Image

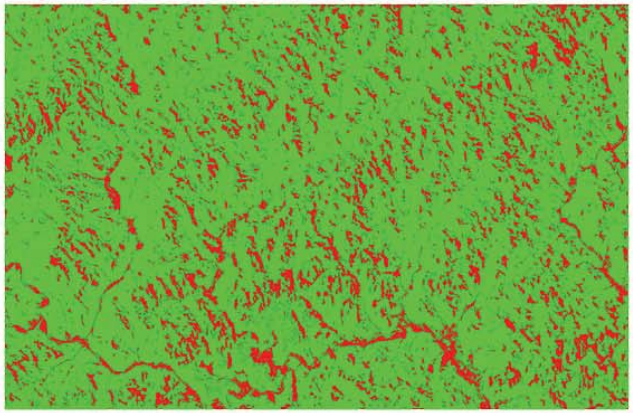
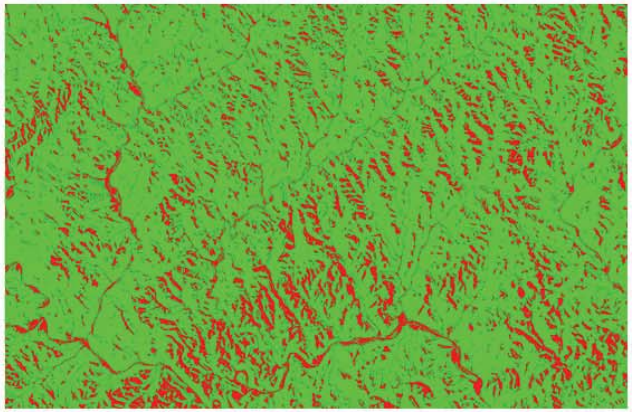
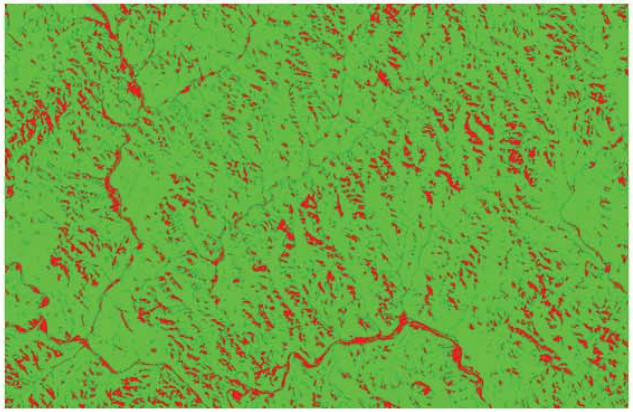


Image Analogy Using Classified Image



Boosted Regression Trees



Originally Classified Image

**Figure 7.4:** Comparison of Binary Classified Images for Pennsylvania Region

# CHAPTER VIII

# DISCUSSIONS & CONCLUSION

This study analyzes the performance of BRT and IA and finding the better technique of the two. This decision is based on a number of important criteria including accuracy, computational speed, and ability to automate the process [13]. Another important criteria is the extent human interpretation and involvement in the process [13]. On this basis, the window selection by the user in IA illustrates that the classification in a way is a reflection of analyst's expectations. The qualitative analysis in Chapter V shows that the results interpreted for each region would change with an individual and is not an undisputed way of providing conclusive results. Classification accuracy is a primary criterion used for analyzing algorithms' performance [13]. It is measured as the percentage of number of pixels correctly classified after using the algorithm [13, 83, 84]. On analyzing all the four cases, BRTs outperform IA. In comparison to conventional classification algorithms, like the maxlik classification method used to classify the image generated from IA, the BRT algorithm has a better accuracy. BRT's better performance in predicting vegetation type is consistent with previous results by Fallah et al, Lawrence et al. and Landenburger et al [85, 86]. In addition to topographic factors, natural disturbances and anthropogenic factors also interrupt the topographic-land cover relation which is a reason why the prediction accuracy of topographic-land cover models can never be absolute [19, 87, 84, 83].

IA is primarily used to develop a "filter" to create effects such as blurring or embossing, improved texture synthesis, super-resolution, texture transfer, artistic filters, etc [42]. This study is first of its kind where IA is being used for land cover classification. This thesis makes use of the boosting algorithm for BRT which draws on the of creating a highly accurate learner by combining several weak learners [31]. This approach has clearly demonstrated that creation of training data set for supervised classification techniques contains subjective

elements which influence training results [88, 89]. The most compelling and useful aspect of this research work is its demonstration of the dominance of various attributes and it can be used for prediction over regions as large as 10720 square kilometers and ability to reproduce land cover maps in resolution of 30 m by 30 m. Previous studies demonstrate that images with spatial resolution of 30 m are sufficient to accurately classify a large variety of landscapes [90, 91, 92, 93, 94, 8].

On comparing the two techniques viz. IA and BRT, IA had the following disadvantage over BRT. The various comparison techniques of the quality of results generated due to IA are not consistent with each other and there no fixed standard method to determine the quality of the image generated. Also, the image B' generated varies with every training pair chosen, so optimization of the window size and region needed to be chosen for training still needs to be determined. On the other hand compared to IA, BRT are computationally very expensive.

When the results of all the regions are analyzed on an aggregated basis, it is seen that these techniques have high performance in regions with more vegetation. The study shows that Hammersley with maximum forested area had a prediction accuracy of 92.86% whilst Los Alamos,the driest of the four regions returned the lowest accuracy of 73%.

The classification accuracy algorithm used is of fundamental importance; land cover classification maps generated in Chapter 7 using different algorithms would not be of much use without knowing the classification accuracies [95, 96]. As evident from the comparison in Chapter 7, BRT is a more accurate and thus reliable method of predicting vegetation cover. Apart from being more accurate, they also return the importance of each attribute used for prediction [12, 28]. Both the IA and BRT methods used here in this study have better prediction accuracies than study by Fallah et al. (2014) which compared efficiency of BRT, Random Forests and Classification and Regression Tree algorithms in land cover classification [12]. The study here when using BRT returned an average accuracy of 80.25% in comparison to Fallah's study with an accuracy of 70%. Even the IA algorithm had a

higher average accuracy of 73% in comparison to most commonly used Classification and Regression Tree method. The comparatively lower performance of IA using Landsat was due to the Maximum Likelihood Classification (Maxlik) algorithm used. Maxlik is a parametric method of classification which is based on the assumption that data comes from a normal Gaussian distribution [97, 98, 99] which was not necessarily true. Unlike, Maxlik method of classification, BRT is a non-parametric method and its classification is not based on any such assumption [99] and thus its improved accuracy.

Clearly from the results of BRT, aspect is the most dominant attribute affecting the land cover followed by elevation in cases where the region has elevation difference. The influence of aspect and elevation can be attributed to the fact that topography influences the vegetation by forming spatially varying micro-climatic zones which further influence the type and growth of vegetation [19]. In mountainous regions as in this study except Hammersley, aspect variations lead to differences in solar radiation, moisture and temperature which in turn leads to various vegetation types and patterns [19, 100, 87, 101]. Aspect's criticality to land cover can be due to factor that it determines the amount of hourly solar radiation received by a surface. Solar radiation affects vegetation growth by affecting soil and air temperatures, and soil moisture conditions [6, 19, 102]. Thus, surfaces receiving more solar radiation experience a more dryer and warmer climatic conditions whereas surfaces facing away from sun have a cooler and moister climates [103, 104, 19]. Slope, too is another important attribute which affects distribution and type of vegetation; steeper slopes affects soil moisture through increased downslope drainage [19, 105, 106]. BRT results show that TI is of less importance in comparison to Aspect, Elevation and Slope in land cover classification. This result is in contrary to the observation by Coblentz and Riitters (2004) which suggested wetness index or TI in our case as a major influential factor limiting vegetation growth in arid regions of the south-western USA and northern Mexico [6].

In comparison to traditional remote sensing methods such as, aerial photography, to develop land cover maps, both the methods presented in this study are inexpensive, faster, and can be extended to regions as large as 10720 square kilometers. The need for this study

is in synergy with past studies which show that large scale data, time-consuming processing, along with integration and interpretation make automated and accurate methods of change land cover mapping highly desirable [107, 108, 109, 109]. If reliable estimates of various forest type over a large area are to be made, there is a need to distinctly identify various forest or vegetation types [110]. Therefore, the limitation of this approach was due to usage of using only binary classes for land cover classification.

Although, numerous studies have been carried out by using variety of ML algorithms for land cover classification, one of the unique features of this study is using Image Analogy for land cover classification which till now has only been used from the point of image processing. This study provides improved prediction accuracies of Boosted Regression Trees and enabling the usage of this technique over large regions in continental United States.

On a concluding note, this work here presents an application using ML techniques to recreate land cover images from satellite images. BRTs look promising in their efficiency and ability to classify large volumes of data and their most important feature of returning importance of attributes. This study can be extended to various other regions in United States and with an accuracy ranging from 73% to 92.86%, this method proves to be reliable. Further scope of this technique can be extended to LULC cover studies to develop climate change model with the inclusion of temporal weather data. This study can be further extended beyond binary classes which would represent a much more vivid scenario in terms of changes in land cover.

# APPENDIX A

# CODES

## *A.1 Image Analogies*

*Navigating to the directory(src) containing the make file for image analogy :*

ms444research@DA102-MS444-1 /cygdrive/c

$ cd Users/ms444research/Dropbox/Hydrological\ Modelling/Felipe\ Dias\ projects/

ms444research@DA102-MS444-1 /cygdrive/c/Users/ms444research/Dropbox/Hydrological Modelling/Felipe Dias projects

$ cd texture_synthesis/princeton_class/image_analogies/src/

ms444research@DA102-MS444-1 /cygdrive/c/Users/ms444research/Dropbox/Hydrological Modelling/Felipe Dias projects/texture_synthesis/princeton_class/image_analogies/src

$ make clean

*Compiling the image analogy files to give an executable file :*

$ make

*Running the image analogy with specified parameters :*

$ ./analogy.exe ../input/Landsat_Wash/A.bmp ../input/

Landsat_Wash/Ap.bmp ../input/Landsat_Wash/B.bmp ../input/Landsat_Wash/Bp.bmp -levels 4 -kappa 30.0 -useAColors

## *A.2 Image Quality Assessment*

clc;

clear all;

% https://www.pantechsolutions.net/blog/matlab-code-for-psnr-and-mse/

*Reading in the original image & coverting it to gray:*

InputImage=rgb2gray(imread('C:\Users\aquarianyashika\Dropbox\Hydrological Modelling
\Felipe Dias projects\texture_synthesis\princeton_class\image_analogies\input\Class_NM
\ClassOriginal.bmp'));

*Reading in the predicted image & coverting it to gray:*

ReconstructedImage=rgb2gray(imread('C:\Users\aquarianyashika\Dropbox\Hydrological Modelling
\Felipe Dias projects\texture_synthesis\princeton_class\image_analogies\input\Class_NM\Bp.bmp'));

n=size(InputImage);

M=n(1);

N=n(2);

*Calculating Mean Square Error:*

MSE = sum(sum((InputImage-ReconstructedImage).$\hat{}$2))/(M*N);

*Calculating PSNR:*

PSNR = 10*log10(256*256/MSE);

%%http://www.mathworks.com/matlabcentral/fileexchange/29500-image-error-measurements/content/
imageQualityIndex.m

*Calculating UIQI:*

uiqi = imageQualityIndex (ReconstructedImage,InputImage);

%%https://ece.uwaterloo.ca/~z70wang/research/ssim/

*Calculating SSIM:*

mssim = ssim( ReconstructedImage,InputImage);

%%http://www4.comp.polyu.edu.hk/~cslzhang/IQA/FSIM/FSIM.htm

*Calculating FSIM:*

$[FSIM, FSIMc]$ = FeatureSIM(InputImage, ReconstructedImage);

```
fprintf('%f\n',MSE)
fprintf('%f\n',PSNR)
fprintf('%f\n',uiqi)
fprintf('%f\n',mssim)
fprintf('%f\n',FSIM)
```

## *A.3 Boosted Regression Trees*

*Loading the gbm package into the workspace :*

> library("gbm", lib.loc="~/R/win-library/3.1")

*Reading in the source file "brt.functions.R" :*

> source.with.encoding('C:/Users/ms444research/Dropbox/JANE_1390_sm_AppendixS3/
brt.functions.R', encoding='UTF-8')

*Reading in the training data set :*

> model.data <−read.csv("C:\\Users\\ms444research\+\Dropbox\\Hydrological Modelling\\Felipe
Dias projects\\BRT\\Penn\\train_data_PA.csv")

*Building a boosted regression tree model :*

> PA.tc5.lr01 <− gbm.step(data=model.data,gbm.x = 2:5,gbm.y = 1,family = "bernoulli",
tree.complexity = 5,learning.rate = 0.01, bag.fraction = 0.5)

*Plotting the dominance of attributes :*

> gbm.plot(PA.tc5.lr01, n.plots=4, write.title = F,plot.layout = c(2,2),cex.lab=1.2)

*Reading in the test data set :*

> eval.data <− read.csv("C:\\Users\\ms444research\+\Dropbox\\Hydrological Modelling\\Felipe
Dias projects\\BRT\\Penn\\test_data_PA.csv")

*Predicting vegetation class using BRT model over the test data set :*

>preds <− predict.gbm(PA.tc5.lr01, eval.data, n.trees=PA.tc5.lr01$gbm.call$best.trees, type="response")

*Exporting the predicted probabilities to a .csv file :*

> write.csv(preds, file = "C:\\Users\\ms444research\\Dropbox\\Hydrological Modelling\\Felipe
Dias projects\\BRT\\Penn\\PA_preds_BRT.csv")

# REFERENCES


[1] "Land Cover Trends Project,".

[2] S. F. Fonji and G. N. Taff, "Using satellite data to monitor land-use land-cover change in North-eastern Latvia," *SpringerPlus* **3**, 61 (2014).

[3] N. Zhao, Y. Yang, and X. Zhou, "Application of geographically weighted regression in estimating the effect of climate and site conditions on vegetation distribution in Haihe Catchment, China," *Plant Ecology* **209**, 349 (2010).

[4] S. Lowe, X. Guo, and D. Henderson, "Landscape spatial structure for predicting suitable habitat: The case of Dalea villosa in Saskatchewan," *Open Journal of Ecology* **2012** (2012).

[5] A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," *Ecological modelling* **135**, 147 (2000).

[6] D. D. Coblentz and K. H. Riitters, "Topographic controls on the regional-scale biodiversity of the south-western USA," *Journal of Biogeography* **31**, 1125 (2004).

[7] N. E. Asselman and H. Middelkoop, "Floodplain sedimentation: quantities, patterns and processes," *Earth Surface Processes and Landforms* **20**, 481 (1995).

[8] O. Rozenstein and A. Karnieli, "Comparison of methods for land-use classification incorporating remote sensing and GIS inputs," *Applied Geography* **31**, 533 (2011).

[9] G. Hegde, J. M. Ahamed, R. Hebbar, and U. Raj, "Urban land cover classification using hyperspectral data," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **1**, 751 (2014).

[10] A. Balmford, R. E. Green, and M. Jenkins, "Measuring the changing state of nature," *Trends in Ecology & Evolution* **18**, 326 (2003).

[11] S. Shirley, Z. Yang, R. Hutchinson, J. Alexander, K. McGarigal, and M. Betts, "Species distribution modelling for the people: unclassified landsat TM imagery predicts bird occurrence at fine resolutions," *Diversity and Distributions* **19**, 855 (2013).

[12] S. Kalbi, A. Fallah, and S. Shataee, "Forest Stand Types Classification Using Tree-Based Algorithms and SPOT-HRG Data," *International Journal of Environmental Resources Research* **1**, 263 (2014).

[13] R. DeFries and J. C.-W. Chan, "Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data," *Remote Sensing of Environment* **74**, 503 (2000).

[14] S. V. Stehman and R. L. Czaplewski, "Introduction to special issue on map accuracy," *Environmental and Ecological Statistics* **10**, 301 (2003).

[15] L. M. de Carvalho, J. G. Clevers, A. K. Skidmore, and S. M. de Jong, "Selection of imagery data and classifiers for mapping Brazilian semideciduous Atlantic forests," *International Journal of Applied Earth Observation and Geoinformation* **5**, 173 (2004).

[16] J. T. Kerr and M. Ostrovsky, "From space to species: ecological applications for remote sensing," *Trends in Ecology & Evolution* **18**, 299 (2003).

[17] W. Turner, S. Spector, N. Gardiner, M. Fladeland, E. Sterling, and M. Steininger, "Remote sensing for biodiversity science and conservation," *Trends in ecology & evolution* **18**, 306 (2003).

[18] S. Franklin and M. Wulder, "Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas," *Progress in Physical Geography* **26**, 173 (2002).

[19] T. K. HAMILTON, "EFFECTS OF TOPOGRAPHY ON THE SPATIAL VARIATION OF LANDCOVER DIVERSITY AND DISTRIBUTION IN A PRAIRIE SANDHILL ECOSYSTEM,".

[20] M. A. Friedl *et al.*, "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of Environment* **83**, 287 (2002).

[21] "A comparison of Machine Learning Algorithms: The Effects of Classification Scheme Detail on Map Accuracy,".

[22] B. Lantz, *Machine Learning with R* (Packt Publishing Ltd, 2013).

[23] Wikipedia, "Machine learning — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 15-August-2014].

[24] R. DeFries and J. Townshend, "NDVI-derived land cover classifications at a global scale," *International Journal of Remote Sensing* **15**, 3567 (1994).

[25] T. Loveland and A. Belward, "The IGBP-DIS global 1km land cover data set, DISCover: first results," *International Journal of Remote Sensing* **18**, 3289 (1997).

[26] S. W. Running, T. R. Loveland, L. L. Pierce, R. Nemani, and E. Hunt, "A remote sensing based vegetation classification logic for global land cover analysis," *Remote sensing of Environment* **51**, 39 (1995).

[27] C. O. Justice, J. Townshend, B. Holben, and e. C. Tucker, "Analysis of the phenology of global vegetation using meteorological satellite data," *International Journal of Remote Sensing* **6**, 1271 (1985).

[28] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology* **77**, 802 (2008).

[29] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology* **88**, 2783 (2007).

[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees* (CRC press, 1984).

[31] R. E. Schapire and Y. Freund, *Boosting: Foundations and algorithms* (MIT press, 2012).

[32] S. Mochizuki and T. Murakami, "Accuracy Comparison of Land Cover Mapping Using the Object- Oriented Image Classification with Machine Learning Algorithms," 2012.

[33] D. Coblentz and P. Keating, "Topographic controls on the distribution of tree islands in the high Andes of south-western Ecuador," *Journal of Biogeography* **35**, 2026 (2008).

[34] J. Franklin, "Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients," *Progress in Physical Geography* **19**, 474 (1995).

[35] A. PéREz, J. F. Mas, A. VELázquEz, and L. VázquEz, "Modelling vegetation diversity types in Mexico based upon topographic features," *Interciencia* **33**, 88 (2008).

[36] M. Austin, "Spatial prediction of species distribution: an interface between ecological theory and statistical modelling," *Ecological modelling* **157**, 101 (2002).

[37] "A planetary-scale platform for environmental data analysis,".

[38] B. L. Foster and T. L. Dickson, "Grassland diversity and productivity: the interplay of resource availability and propagule pools," *Ecology* **85**, 1541 (2004).

[39] D. Tilman *et al.*, "Productivity and sustainability influenced by biodiversity in grassland ecosystems," *Nature* **379**, 718 (1996).

[40] D. Tilman, P. B. Reich, J. Knops, D. Wedin, T. Mielke, and C. Lehman, "Diversity and productivity in a long-term grassland experiment," *Science* **294**, 843 (2001).

[41] A. Hector *et al.*, "Plant diversity and productivity experiments in European grasslands," *science* **286**, 1123 (1999).

[42] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* pp. 327–340 ACM 2001.

[43] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence* **14**, 1612 (1999).

[44] "Chris Tralie : Image Analogies,".

[45] "Image Quality Assessment,".

[46] D. D. C. S. Yusra A. Y. Al-Najjar, "Comparison of Image Quality Assessment: PSNR, HVS, SSIM, UIQI,".

[47] Z. Wang and A. C. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE* **9**, 81 (2002).

[48] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on* volume 2 pp. 1398–1402 Ieee 2003.

[49] Wikipedia, "Structural similarity — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 20-August-2014].

[50] S. Jun, "Boosted regression trees and random forests," *Statistical Consulting Report for Michael Melnychuck. University of British Columbia* (2013).

[51] J. M. Allad, "An Application of Gradient Boosted Regression Trees and Random Forests to Prospect Direct Marketing Response Modeling,".

[52] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques* (Morgan Kaufmann, 2005).

[53] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics* , 1189 (2001).

[54] G. De'Ath, "Boosted trees for ecological modeling and prediction," *Ecology* **88**, 243 (2007).

[55] R. E. Schapire, "The boosting approach to machine learning: An overview,", in *Nonlinear estimation and classification* pp. 149–171 Springer 2003.

[56] Google, "Google Earth Engine,".

[57] Wikipedia, "GRASS GIS — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 28-January-2015].

[58] Wikipedia, "R (programming language) — Wikipedia, The Free Encyclopedia," 2015, [Online; accessed 28-January-2015].

[59] A. Morton, "UTM Grid Zones of the World,".

[60] "National Forests of Alabama,".

[61] "Regional Climate Maps - Twelve Month Extreme Average Temperature,".

[62] "Regional Climate Maps - Twelve Month Extreme Minimum Temperature,".

[63] "Regional Climate Maps - Twelve Month Extreme Maximum Temperature,".

[64] "Regional Climate Maps - Twelve Month Total Precipitation,".

[65] "USGS National Elevation Dataset 1/3 arc-second,".

[66] Wikipedia, "William B. Bankhead National Forest — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 15-December-2014].

[67] Wikipedia, "Alpine Lakes Wilderness — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 15-December-2014].

[68] "Andisols Map,".

[69] "Inceptisols Map,".

[70] "Land Cover Map for the Eastern Jemez Region,".

[71] "Dominant Soil Orders,".

[72] Wikipedia, "Baraboo, Wisconsin — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 17-December-2014].

[73] "Regional Property Analysis: Sauk Prairie Recreation Area,".

[74] Wikipedia, "Prescott, Arizona — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[75] Wikipedia, "Waynesville, North Carolina — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[76] Wikipedia, "Reno, Nevada — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[77] Wikipedia, "Cincinnati — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[78] Wikipedia, "Aspen, Colorado — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[79] "Rocky Mountain,".

[80] "South Yolla Bolly (Mt Linn),".

[81] Wikipedia, "Rio Grande National Forest — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 19-December-2014].

[82] Wikipedia, "Hammersley Wild Area — Wikipedia, The Free Encyclopedia," 2014, [Online; accessed 20-December-2014].

[83] N. Oreskes *et al.*, "Verification, validation, and confirmation of numerical models in the earth sciences," *Science* **263**, 641 (1994).

[84] J. Franklin, P. McCullough, and C. Gray, "Terrain variables used for predictive mapping of vegetation communities in Southern California," *Terrain analysis: principles and applications/edited by John P. Wilson, John C. Gallant* (2000).

[85] R. Lawrence, A. Bunn, S. Powell, and M. Zambon, "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis," *Remote sensing of environment* **90**, 331 (2004).

[86] L. Landenburger, R. L. Lawrence, S. Podruzny, and C. C. Schwartz, "Mapping Regional Distribution of a Single Tree Species: Whitebark Pine in the Greater Yellowstone Ecosystem," *Sensors* **8**, 4983 (2008).

[87] B. Hoersch, G. Braun, and U. Schmidt, "Relation between landform and vegetation in alpine regions of Wallis, Switzerland. A multiscale remote sensing and GIS approach," *Computers, Environment and Urban Systems* **26**, 113 (2002).

[88] D. McIver and M. Friedl, "Using prior probabilities in decision-tree classification of remotely sensed data," *Remote Sensing of Environment* **81**, 253 (2002).

[89] G. Foody, M. McCulloch, and W. Yates, "The effect of training set size and composition on artificial neural network classification," *International Journal of Remote Sensing* **16**, 1707 (1995).

[90] M. Alrababah and M. Alhamad, "Land use/cover classification of arid and semi-arid Mediterranean landscapes using Landsat ETM," *International journal of remote sensing* **27**, 2703 (2006).

[91] N. Koutsias and M. Karteris, "Classification analyses of vegetation for delineating forest fire fuel complexes in a Mediterranean test site using satellite remote sensing and GIS," *International Journal of Remote Sensing* **24**, 3093 (2003).

[92] R. Manandhar, I. O. Odeh, and T. Ancev, "Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement," *Remote Sensing* **1**, 330 (2009).

[93] S. A. Sader, D. Ahl, and W.-S. Liou, "Accuracy of Landsat-TM and GIS rule-based methods for forest wetland classification in Maine," *Remote Sensing of Environment* **53**, 133 (1995).

[94] J. J. Schulz, L. Cayuela, C. Echeverria, J. Salas, and J. M. R. Benayas, "Monitoring land cover change of the dryland forest landscape of Central Chile (1975–2008)," *Applied Geography* **30**, 436 (2010).

[95] N. Baatuuwie and L. Van Leeuwen, "Evaluation of three classifiers in mapping forest stand types using medium resolution imagery: A case study in the Offinso Forest District, Ghana," *African Journal of Environmental Science and Technology* **5**, 25 (2011).

[96] L. Yang, S. V. Stehman, J. H. Smith, and J. D. Wickham, "Thematic accuracy of MRLC land cover for the eastern United States," *Remote sensing of Environment* **76**, 418 (2001).

[97] B. W. Szuster, Q. Chen, and M. Borger, "A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones," *Applied Geography* **31**, 525 (2011).

[98] J. R. Jensen and K. Lulla, "Introductory digital image processing: a remote sensing perspective," (1987).

[99] Wikipedia, "Parametric statistics — Wikipedia, The Free Encyclopedia," 2015, [Online; accessed 19-March-2015].

[100] T. R. Oke, *Boundary layer climates* (Routledge, 2002).

[101] Y. Deng, X. Chen, E. Chuvieco, T. Warner, and J. P. Wilson, "Multi-scale linkages between topographic attributes and vegetation indices in a mountainous landscape," *Remote Sensing of Environment* **111**, 122 (2007).

[102] J. Bennie, M. O. Hill, R. Baxter, and B. Huntley, "Influence of slope and aspect on long-term vegetation change in British chalk grasslands," *Journal of Ecology* **94**, 355 (2006).

[103] P. Miller and D. Poole, "The influence of annual precipitation, topography, and vegetative cover on soil moisture and summer drought in southern California," *Oecologia* **56**, 385 (1983).

[104] C. D. Ahrens, *Essentials of meteorology: an invitation to the atmosphere* (Cengage Learning, 2011).

[105] F. T. Maestre, J. Cortina, S. Bautista, J. Bellot, and R. Vallejo, "Small-scale environmental heterogeneity and spatiotemporal dynamics of seedling establishment in a semiarid degraded ecosystem," *Ecosystems* **6**, 630 (2003).

[106] D. E. Koenig, *The effects of dune stabilization on the spatiotemporal distribution of soil moisture resources, Northern Great Plains, Canada*, PhD thesis Lethbridge, Alta.: University of Lethbridge, Dept. of Geography, c2012 2012.

[107] M. C. Hansen, J. C.-W. Chan, J. Pagis, R. DeFries, and D. Luo, "Long term change detection using continuous fields of tree cover from 8km AVHRR data for the years 1982-2000," *Analysis of Multi-Temporal Remote Sensing Images* **2**, 363 (2002).

[108] R. Aspinall, "A land-cover data infrastructure for measurement, modeling, and analysis of land-cover change dynamics," *Photogrammetric engineering and remote sensing* **68**, 1101 (2002).

[109] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts, "Mapping land-cover modifications over large areas: A comparison of machine learning algorithms," *Remote Sensing of Environment* **112**, 2272 (2008).

[110] S. E. Franklin, *Remote sensing for sustainable forest management* (CRC Press, 2001).